Bayesian Multisensory Perception

Timothy M. Hospedales

Doctor of Philosophy

Institute of Perception, Action and Behaviour
School of Informatics
University of Edinburgh

2008

Abstract

A key goal for humans and artificial intelligence systems is to develop an accurate and unified picture of the outside world based on the data from any sense(s) that may be available. The availability of multiple senses presents the perceptual system with new opportunities to fulfil this goal, but exploiting these opportunities first requires the solution of two related tasks. The first is how to make the best use of any redundant information from the sensors to produce the most accurate percept of the state of the world. The second is how to interpret the relationship between observations in each modality; for example, the correspondence problem of whether or not they originate from the same source.

This thesis investigates these questions using ideal Bayesian observers as the underlying theoretical approach. In particular, the latter correspondence task is treated as a problem of Bayesian model selection or structure inference in Bayesian networks. This approach provides a unified and principled way of representing and understanding the perceptual problems faced by humans and machines and their commonality.

In the domain of machine intelligence, we exploit the developed theory for practical benefit, developing a model to represent audio-visual correlations. Unsupervised learning in this model provides automatic calibration and user appearance learning, without human intervention. Inference in the model involves explicit reasoning about the association between latent sources and observations. This provides audio-visual tracking through occlusion with improved accuracy compared to standard techniques. It also provides detection, verification and speech segmentation, ultimately allowing the machine to understand *"who said what, where?"* in multi-party conversations.

In the domain of human neuroscience, we show how a variety of recent results in multimodal perception can be understood as the consequence of probabilistic reasoning about the causal structure of multimodal observations. We show this for a localisation task in audio-visual psychophysics, which is very similar to the task solved by our machine learning system. We also use the same theory to understand results from experiments in the completely different paradigm of oddity detection using visual and haptic modalities. These results begin to suggest that the human perceptual system performs – or at least approximates – sophisticated probabilistic reasoning about the causal structure of observations under the hood.

Acknowledgements

I would like to thank Sethu Vijayakumar, my supervisor, for his inspiration, support and guidance throughout my time in Edinburgh. I would also like to thank Marc Toussaint and Mark van Rossum, for providing their insight and many helpful discussions.

I must also thank Oliver Williams and Andrew Blake at Microsoft Research for giving me the invaluable opportunity to visit and work with them.

I am grateful also for the moral support and interesting discussions provided by all my follow students in the SLMC research group, the Neuroinformatics DTC and the ANC and IPAB institutes.

Finally, thank you to all my family for their constant support, and thank you to Julieta for making my life brighter and happier in these past years.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Timothy M. Hospedales*)

To my parents, James and Shelley

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A key goal for humans and artificial intelligence systems is to develop an accurate and unified picture of the outside world based on the data from any sense(s) that may be available. The availability of multiple senses presents the perceptual system with new opportunities to fulfil this goal, but exploiting these opportunities first requires the solution of some interesting computational problems.

In the event that each sensor carries redundant information about the uncertain state of a source, multiple sensors allow the state to be inferred more precisely. For example, an aircraft control system may compute the best estimate of an aeroplane's location given the readings from two different radar towers. As another example, consider searching for your dog, which has run off in the forest during a walk. You may navigate toward some combination of moving leaves on the horizon and the dog's bark. In both examples, using the two independent observations can improve the accuracy of localization. To model this type of task in a graphical model formalism [Bishop, 2006a], we can use a graph like Figure 1.1. Here an object's unknown state (e.g., the dog or aeroplane's location) produces some observations x_i (e.g., radar returns or moving leaves and barks) which are probabilistically related to the true state y. The computations required for optimal (i.e., minimum variance) estimation of the state given the observations in such models are relatively straightforward [Clark and Yuille, 1990], and we term this process sensor *fusion*.

Some other quantities of interest are defined by the *relation* between observations in multiple modalities, and may not correspond directly to physical properties of the world (such as location). For example, when searching for your lost dog, using both auditory and visual information may help you localise it more accurately. However, consider the scenario in which you may hear a bark, but recognise it as coming from a

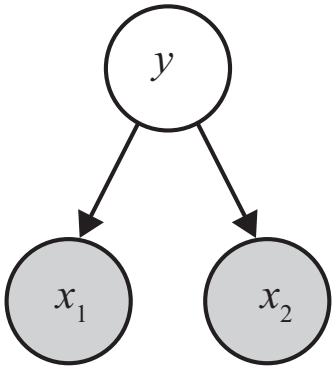


Figure 1.1: Graphical model of the classical sensor fusion scenario, in which observations x_i, produced by a source of unknown state y, are used to infer the state y.

totally different direction than the one your dog ran off in, or being in a totally different tone than your dog's bark. In this case, you may decide that the sound must be caused by some other animal and therefore, having no relation to your lost dog, should be discounted in your search. This example includes uncertain causal structure, where the causality of your observation (did this observation come from my pet, or some other pet?) is uncertain and also to be computed. So the significance of a particular observation depends on its relation to its source and other observations, but these relations may themselves be uncertain. The questions of the unknown state which may be responsible for the observations, and their causal relation to the state, are conditionally dependent and as such cannot be considered independently for optimal inference.

In some situations, this relation between observations may be of key independent interest (the actual content of the observations may even be only secondary!). For example, in a meeting, while it is important to understand directly observed audio-visual quantities such as *what was said* and *who was there*, it may be equally important to understand the *relation* between the audio-visual observations, ***who said what?***

In this thesis, we investigate such problems in multisensor perception in which both state and relational structure are unknown. We term these two related problems *state* and *structure* inference respectively. A key novel step in this work is the casting of such tasks with unknown causal structure as problems of *Bayesian model selection*. This allows us to construct formal probabilistic models of these tasks using Bayesian networks with unknown latent state variables and causal structure (conditional dependencies). Using this common theoretical ground, we are able to understand the commonality in problems of multisensory perception spanning both the domains of artificial intelligence / machine learning and human neuroscience. This enables us to

investigate and contribute to both domains. We now discuss the application of these concepts to machine perception problems in some more detail.

1.1 Machine Learning

Optimal fusion of redundant multisensor observations can be applied to build better machine perception applications. In principle, multisensor fusion is useful to an agent because more precise inferences about the world can be drawn given multiple observations with independent noise. For example, in Figure 1.1, assume our prior knowledge about y is given by $y \sim \mathcal{N}(0, p_0)$, and assume observations x_i are Gaussian $x_i \sim \mathcal{N}(y, p_i)$ with precision p_i. (Note that we use precisions rather than variances here[1].) Then, the posterior distribution $p(y|x_1,,x_2)$ is Gaussian $\mathcal{N}(y|\mu_{y|x_1,x_2}, p_{y|x_1,x_2})$ with statistics[2]:

$$\mu_{y|x_1,x_2} = \frac{p_1 x_h + p_2 x_v}{p_0 + p_1 + p_2}, \tag{1.1}$$

$$p_{y|x_1,x_2} = p_0 + p_1 + p_2. \tag{1.2}$$

The increased precision of the posterior distribution $p_{y|x_1,x_2}$ is clear, as it is the sum of the precision of the individual observations (eq. (1.2)). The best estimate $\hat{y}$ can simply be taken to be the mean $\mu_{y|x_1,x_2}$ of the posterior distribution (eq. (1.1)).

Applying statistical learning techniques to machine perception problems, fusion of multiple modalities or features is a common technique to improve performance. In speech recognition, for example, visual lip features have been fused with audio data to improve recognition performance [Nefian et al., 2002]. In tracking, performance has been enhanced by fusion of color, texture and edge features within video [Serby et al., 2004], as well as fusing entirely separate audio and video modalities [Beal et al., 2003, Perez et al., 2004, Hershey and Movellan, 1999, Chen and Rui, 2004].

All these studies have generally considered cases in which the observations are known to be generated from the same latent source (Figure 1.1), and the task is to

[1] In this thesis, we will use the notation $x \sim \mathcal{N}(\mu, p)$ or $\mathcal{N}(x|\mu, p)$ equivalently to denote a random variable x distributed normally with mean μ and precision p. Precisions are used throughout, except in the context of the psychophysics model in Section 4.3, where we use variances to maintain consistency with the experiment being modelled.

[2] See Appendix A.1.2.3 for details and derivation. Note that in our notation, for a distribution such as $\mathcal{N}(y|\mu_{y|x_1,x_2}, p_{y|x_1,x_2})$, the sufficient statistics $\mu_{y|x_1,x_2}$ and $p_{y|x_1,x_2}$ are written with subscripts x_1, x_2 to indicate, for clarity, the data which the distribution is conditioned on.

make the best estimate of the latent source state by fusing the observations — we will call models assuming such a fused structure *pure fusion* or *classical sensor fusion* models. However, as we have seen, in many real world situations, any given pair of observations are unlikely to have originated from the same latent source. The perceptual system is therefore faced with a type of correspondence (or binding) problem ([Treisman, 1996]). The more general task in multisensor perception is therefore to infer the *association* between observations and any latent states of interest as well as the latent state itself. Inference of the latent state may require fusion (integration) or fission (segregation) depending on the association.

This type of problem has been of long standing interest in the radar community where it is known as *data association* [Bar-Shalom and Tse, 1975, Bar-Shalom et al., 2005]. Here, the association decisions might, for example, be made between a pool of candidate radar detections and existing aircraft tracks before the tracks are be updated on the basis of the new observations. However, popular methods in this domain [Bar-Shalom and Tse, 1975, Bar-Shalom et al., 2005] have tended to be heuristic heavy due to the strict real time requirements coupled with typically high dimensional, large data sets, with some notable exceptions [Stone et al., 1999, Vermaak et al., 2005].

In a probabilistic modelling context, data association is an example of a structure inference or model selection problem. Here, the potential existence of a causal connection between a given pair of latent and observed variables is itself an unknown. Early studies of this type of uncertain structure problem by the probabilistic modelling community described efficient inference for some classes of network using Bayesian multinets [Geiger and Heckerman, 1996]. The Bayesian multinet approach has been applied, for example, to infer the (time varying) connectivity structure in Markov chains [Bilmes, 2000]. If potential conditional indepedencies are not known a priori, they can themselves be discovered in data using Context Specific Independence [Boutilier et al., 1996]. All this is in contrast to *learning* a *fixed* Bayesian network structure from large data sets, which is also topical [Silva and Scheines, 2006, Mansinghka et al., 2006].

Inspired by radar/sonar data association algorithms [Fortmann et al., 1983], some machine learning for computer vision studies have begun to consider this issue [Rasmussen and Hager, 2001]. Nevertheless, computer vision studies have tended to see data association as a nuisance variable: a prerequisite for correct fusion in a multisensor and/or multi-target context, but otherwise uninteresting, and to be in-

tegrated out as quickly and efficiently as possible. In contrast, we will argue that for many applications, the association is itself a useful output worthy of careful explicit modelling and consideration. Data association can be of intrinsic interest for understanding complex semantic structure in the data. This is clearly the case in problems of audio-visual perception, where the association represents *who said what*. For example, typical meeting room goals for a human or automatic transcription machine [Hain et al., 2005] might include understanding speech and identifying the participants. However, without explicitly computing the association between audio and visual observations and latent sources, such an agent might have a notion of 'who was there' and 'what was said', but not '*who said what*' – a *relational* concept, specifying the existence of causal connections between different variables which are critical to the meeting understanding problem. Some recent studies have included computation of speaker association in audio-visual tracking for meeting analysis using particle filters [Gatica-Perez et al., 2007, Checka et al., 2004]. For computing data association, an alternative approach to structure inference in explicit parametric models is based on computing and thresholding the mutual information between modalities [Slaney and Covell, 2000, Fisher and Darrell, 2004]. However, this has the drawback of being purely a method to estimate association without a principled framework for simultaneous inference of other quantities of interest such as tracking [Gatica-Perez et al., 2007] or recognition [Nefian et al., 2002] which parametric models can provide.

1.2 Theoretical Neuroscience

Beyond the building of machine applications, Bayesian probabilistic modelling of perceptual tasks is as an elegant and successful approach to understanding many aspects of human perception. In psychophysics, this is frequently called *ideal observer* modelling. It has seen extensive application in visual perception (e.g., [Kersten et al., 2004, Yuille and Kersten, 2006]); in understanding how prior belief is combined with observations (e.g., in perception [Kersten et al., 2004] and sensorimotor control [Kording and Wolpert, 2004a]); and in understanding how multisensory information is combined into a unified percept [Ernst and Bulthoff, 2004]. In the case of multisensory perception – the topic of this thesis – standard probabilistic models for sensor fusion (Figure 1.1) and the resultant equations for inference (eqs. (1.1) and (1.2)) turn out to be a good explanation of human perception for many tasks and pairs

of senses.

For example, in the audio-visual domain, disparate per-scenario theories were previously posed for multisensory spatial and frequency perception. Visual capture was posed to explain the observed visual dominance in spatial judgement scenarios such as the ventriloquist effect [Alais and Burr, 2004] and auditory-capture to explain the auditory dominance observed in frequency judgement tasks [Shams et al., 2000, Recanzone, 2003]. However, these can both be understood [Alais and Burr, 2004, Witten and Knudsen, 2005] as special cases of a single principle of optimal sensor fusion, for scenarios in which visual observations have much higher precision (spatial measurement) and those in which audio observations have much higher precision (frequency measurement), respectively. In these cases, because of the estimation by weighted mean predicted by classical sensor fusion eq. (1.1), the higher precision modality would appear to dominate when modalities are combined. Under experimental intervention to manipulate the precision of the modalities, the final percept does indeed vary smoothly as a function of the individual observations and their precisions; thus optimal fusion eq. (1.1), rather than simple one-way dominance, is revealed to be the underlying principle implemented by the brain [Ernst and Bulthoff, 2004, Witten and Knudsen, 2005].

Recent studies in human multisensory perception have highlighted a variety of pairs of senses and tasks for which human performance in multisensory perception tasks is near Bayes optimal (i.e., conforming with eqs. (1.1) and (1.2)). Examples include visual-haptic size perception [Ernst and Banks, 2002, Gepshtein and Banks, 2003], visual-proprioceptive hand localization [van Beers et al., 2002] and audio-visual spatial localization [Alais and Burr, 2004]. Different cues within a given sensory modality can also independently provide information about a given stimulus source and hence, provide an additional opportunity for sensor fusion. Within vision, texture and motion cues to depth [Jacobs, 1999] and texture and stereo cues to slant [Hillis et al., 2004] appear to be combined according to optimal sensor fusion principles.

In some recent studies, classical sensor fusion as a theory for human multisensory perception has broken down as an explanation of the data [Hillis et al., 2002, Hairston et al., 2003, Wallace et al., 2004, Roach et al., 2006, Recanzone, 2003]. Interestingly, these experiments share the common feature that they have presented multisensory stimuli which are sufficiently discrepant that it is no longer reasonable to assume that the causal relation between the multisensory observations is fixed (as in

Figure 1.1). In these cases, we should perhaps not be surprised that the classical sensor fusion models fail to explain the data well. If the brain is indeed performing Bayesian inference about object state given its observations, it should also infer distributions over the causal structure if this is also uncertain. In Chapter 4, we will investigate whether structure inference can provide an explanation for some of these experiments, thereby maintaining Bayesian inference or ideal observers as a general theory for human multisensory perception. Specifically, we consider two sets of experiments: investigating localization in the audio-visual [Hairston et al., 2003, Wallace et al., 2004] domains, and oddity detection in the visual and visual-haptic [Hillis et al., 2002] domains.

1.3 Thesis Outline

This theses addresses theoretical and applied questions in multisensory perception, particularly those where the correspondence between observations and latent sources is not entirely known. **Chapter 2** introduces the use of Bayesian networks and Bayesian model selection as a common theoretical framework for modelling these types of problems. Solutions to toy problems are illustrated to give insight into the modelling framework before more complicated, realistic applications are considered in later chapters. Applying this basic theory, **Chapter 3** describes a large scale machine learning system developed for machine understanding of audio-visual scenes. The theoretical and modelling details for this application are described in detail in Chapter 3, Section 3.2. Extensive evaluation, highlighting the benefits of the Bayesian structure inference approach to audio-visual machine learning is conducted in Chapter 3, Section 3.3. Our theoretical approach can also be applied as a model of human multisensory perception. In **Chapter 4**, we show how a variety of previously unexplained recent results in neuroscience can be understood a consequence of the perceptual system performing probabilistic inference and model selection in computing the percept of the world. In Chapter 4, Section 4.2 we model experiments on human audio-visual localization, which correspond to the task which we build a machine learning system to perform in Chapter 3. In Chapter 4, Section 4.3 we model experiments on human visual-haptic oddity perception, in which the model selection approach developed in this thesis turns out to be critical to explain the results. Finally, **Chapter 5** concludes the thesis, discussing the commonality of the problems we address in artificial intelligence and neuroscience, discussing avenues for future work and summarising our contributions.

Chapter 2

Theoretical Framework for Multisensory Perception

In this chapter, we introduce the probabilistic foundations of multisensor perception problems where the data association is unknown. To clearly illustrate the method and benefits of this approach, we describe the task and generic modelling framework by way of a series of toy models for single (Section 2.1) and multiple (Section 2.2) latent sources.

2.1 Modelling a Single Source

In order to formalize the perceptual problem of combining information from multiple sensory modalities to obtain an accurate, unified percept of the world, we use a probabilistic generative modelling framework. The task of perception can be abstracted to one of performing *inference* in the generative model , where 'latent' quantities of interest (e.g., location of a person) are inferred on the basis of sensor observations. Figure 2.1 represents this situation when the association – or structure - between the latent source state y and observations x_i is unambiguously known.

In the event that the data association (e.g., who said what) is not known a priori, we need a framework capable of inferring both the state and the association. We can frame such inference as a model selection (or structure inference) problem as schematically represented by the graphical models in Figure 2.2. Here, observations in two *different modalities* $D = \{x_1, x_2\}$ are potentially generated from a *single* source with latent state l (Figure 2.2(a)). Under this generative model, the source state is assumed drawn independently along with binary visibility/occlusion variables $(\mathrm{M}_1, \mathrm{M}_2)$ in each modality.

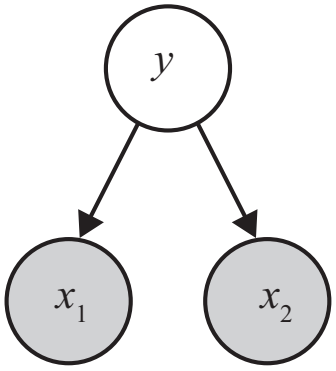


Figure 2.1: Graphical model of the classical sensor fusion scenario, in which observations x_i, produced by a source of unknown state y, are used to infer the state y. The association between the source state and observations is assumed to be unambiguous.

Subsequently, the observations are generated with x_i being dependent on l if $\mathrm{M}_i = 1$ or on a background distribution if $\mathrm{M}_i = 0$. Alternately, all the structure options could be explicitly enumerated into four separate models (Figure 2.2(b)).

Perceptual inference then consists of computing the posterior over the latent state and the generating model (either as specified by the two binary structure variables M_i or a single model index variable) given the observations. An observation in modality i is perceived as being associated with (having originated from) the latent source of interest with probability $p(\mathrm{M}_i = 1|D)$. This will be large if the observation is likely under the foreground distribution (i.e. correlated with the prior and other observations) and small if it is better explained by the background distribution.

2.1.1 An Illustrative Example

To illustrate with a toy but concrete example, consider the problem of inferring a single dimensional latent state l representing a location on the basis of two point observations in separate modalities. For the purpose of this illustration, let the latent location be governed by an informative Gaussian[1] prior $l \sim \mathcal{N}(l|0, p_l)$ with the binomial visibility variables having prior probability $p(\mathrm{M}_i = 1) = \pi_i$. If the state is observed by sensor i ($\mathrm{M}_i = 1$), then the observation in that modality is generated with precision p_i, such that $x_i \sim \mathcal{N}(x_i|l, p_i)$. Alternately, if the state is not observed by the sensor, its observation is generated by the background distribution $\mathcal{N}(x_i|0, p_b)$, which tends toward

[1]The assumption of a one dimensional Gaussian prior and likelihoods is to facilitate illustrative analytical solutions; this is not in general a restriction of our framework as can be seen in Section 3.2.

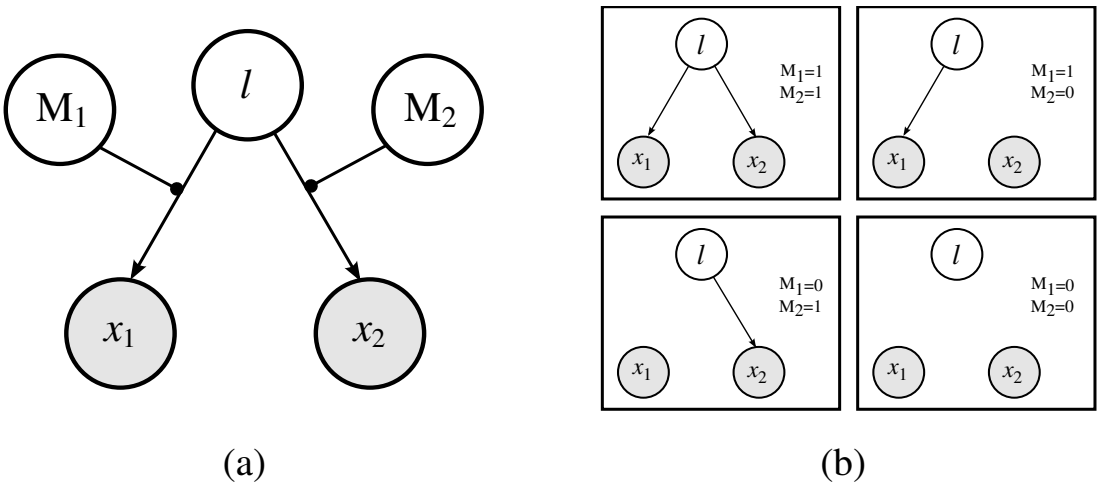


Figure 2.2: Graphical models to describe "unreliable generation" of multimodal observations from a single source. (a) Variable structure interpretation. (b) Variable model interpretation.

un-informativeness with precision $p_b \to 0$. The joint probability can then be written:

$$\begin{aligned} p(D,l,\mathrm{M}_1,\mathrm{M}_2) &= \mathcal{N}(x_1|l,p_1)^{\mathrm{M}_1}\mathcal{N}(x_1|0,p_b)^{(1-\mathrm{M}_1)}\mathcal{N}(x_2|l,p_2)^{\mathrm{M}_2} \\ &\quad \cdot\mathcal{N}(x_2|0,p_b)^{(1-\mathrm{M}_2)}\mathcal{N}(l|0,p_l)p(\mathrm{M}_1)p(\mathrm{M}_2). \end{aligned} \tag{2.1}$$

If we are purely interested in computing the posterior over latent state l, we integrate over models or structure variables: $\sum_{\mathrm{M}_1,\mathrm{M}_2} p(D,l,\mathrm{M})$. For the higher level task of inferring the cause or association of observations, we integrate over the state to compute the posterior model probability, benefiting from the automatic complexity control induced by Bayesian Occam's razor [MacKay, 2003]. We define for brevity the indicators m_i and $\overline{m}_i$, representing the cases of association $(\mathrm{M}_i = 1)$ and disassociation $(\mathrm{M}_i = 0)$ respectively. Then, based on eq. (2.1), we can write down the data likelihoods as in eqs. (2.2)-(2.4)[2]:

$$p(D|\overline{m}_1,\overline{m}_2) \propto \mathcal{N}(x_1|0,p_b)\mathcal{N}(x_2|0,p_b), \tag{2.2}$$

$$p(D|m_1,\overline{m}_2) \propto \exp\left(-\frac{1}{2}x_1^2 p_1 p_l/(p_1+p_l)\right)\mathcal{N}(x_2|0,p_b), \tag{2.3}$$

$$p(D|m_1,m_2) \propto \exp\left[-\frac{1}{2}\frac{x_1^2 p_1(p_2+p_l) - 2x_1x_2p_1p_2 + x_2^2p_2(p_1+p_l)}{p_1+p_2+p_l}\right]. \tag{2.4}$$

The ***structure*** *posterior* $p(\mathrm{M}|D) = p(D|\mathrm{M})p(\mathrm{M})/p(D)$ is dependent on the relative data likelihood under the background and the marginal foreground distribution. For

[2]See Appendix A.1.2.4 for derivation and details

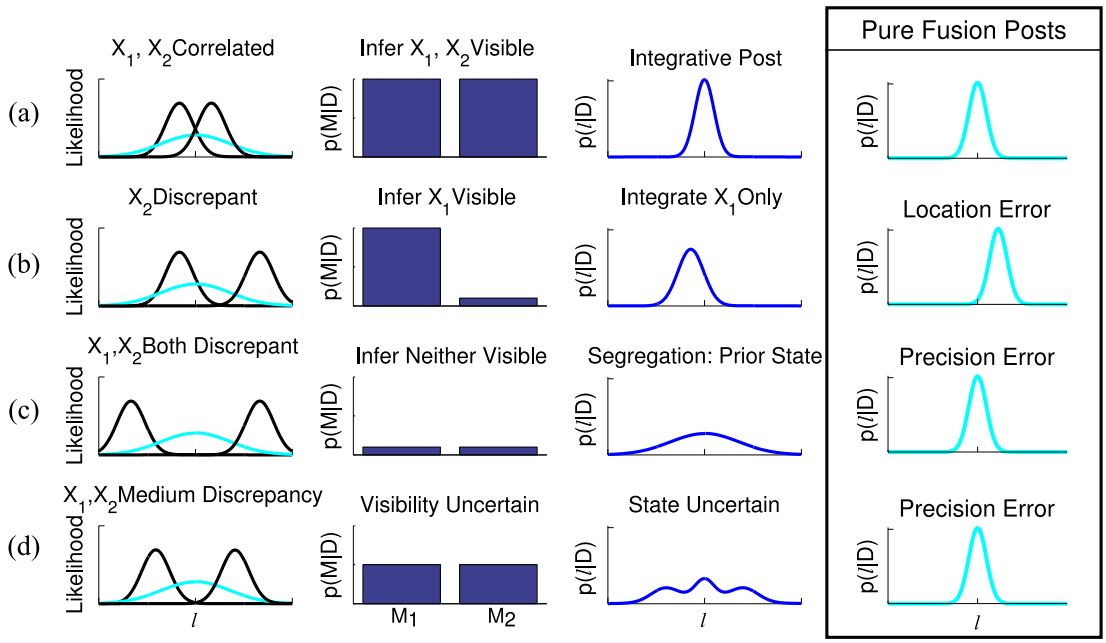


Figure 2.3: Schematic of data association inference given multimodal observations, x_i. Likelihoods of the observations in each of two modalities are in black, prior is in grey. Observations (a) x_1, x_2 strongly correlated, (b) x_2 strongly discrepant, (c) x_1, x_2 both strongly discrepant, (d) x_1, x_2 both moderately discrepant.

example, the posterior of the completely disassociated model eq. (2.2) depends on the background distribution likelihoods, which only vary weakly with the data due to their low precision. This posterior therefore, primarily depends on the data via the normalisation constant. In contrast, the posterior of the fully associated model eq. (2.4) depends on the three way agreement between the observations and the prior. The model structure inference computed using eqs. (2.2)-(2.4) is plotted as $p(\mathrm{M}|D)$ in Figure 2.3 for various illustrative cases.

The convenient form of the structure posterior $p(M|D)$ (eqs. (2.2)-(2.4)), can be used to easily compute the ***latent state*** *'location'* posterior $p(l|D)$ as the following mixture of Gaussian density:

$$p(l|D) \quad = \quad \sum_{\mathrm{M}_1,\mathrm{M}_2} p(l|\mathrm{M}_1,\mathrm{M}_2,D)p(\mathrm{M}_1,\mathrm{M}_2|D), \tag{2.5}$$

where the model conditional location posterior term $p(l|\mathrm{M}_1,\mathrm{M}_2,D)$ is computed from classical sensor fusion equations (eqs. (1.1) and (1.2)). Figure 2.3 also plots this inferred latent location posterior $p(l|D)$, which can be contrasted with the pure fusion models (refer Figure 2.3(box)) estimates as follows:

- [Figure 2.3(a)] Observations and the prior are all strongly correlated: Both observations are inferred to be associated with latent source. The location posterior is approximately Gaussian with $p(l|x_1,x_2) \approx \mathcal{N}(l|\hat{l}, p_{l|x})$ where $p_{l|x} = p_1 + p_2 + p_l$ and $\hat{l} = \frac{p_1 x_1 + p_2 x_2}{p_{l|x}}$. This matches the pure fusion estimates.

- [Figure 2.3(b)] Observation x_2 is strongly discrepant with x_1 and the prior: Sensor 2 is inferred to be occluded. The resultant approximately Gaussian location posterior fuses only x_1 and the prior; $p_{l|x} = p_1 + p_l$, $\hat{l} = \frac{p_1 x_1}{p_{l|x}}$. Pure fusion posterior modes can be displaced arbitrarily far away from the actual source as a consequence of fusing the unrelated sensor (Figure 2.3(b)(box)).

- [Figure 2.3(c)] Observations x_1 and x_2 are strongly discrepant with each other and the prior: Both observations are inferred to be unrelated to actual source (both sensors occluded), in which case the posterior over the latent state reverts to the prior $p_{l|x} = p_l$, $\hat{l} = 0$. In the pure fusion models, posterior distributions could indicate dramatically inappropriate over-confidence (Figure 2.3(c)(box)).

- [Figure 2.3(d)] Correlation between the observations and the prior is only moderate: The posteriors over structural visibility variables are highly uncertain. The location posterior is a (potentially quad-modal) mixture of Gaussians corresponding to the four possible models. Again, the pure fusion model displays inappropriate over-confidence over location (Figure 2.3(d)(box)).

In real world scenarios, occlusion, sensor failure, or other cause for meaningless observation is almost always possible. In these cases, assuming a typical pure fusion model (equivalent to constraining $\mathrm{M}_1 = \mathrm{M}_2 = 1$) can result in dramatically inappropriate inference (as illustrated in Figure 2.3(box) and the explanation above). Examples of these types of effect in real data will be illustrated in Section 3.3.1. The biggest benefit of our approach, however, will be evident in real world applications where meaningful sources and observations result in data association (inferred through the structure posterior) having important relational consequence rather than merely ensuring robust tracking.

2.1.2 Incorporating Temporal Dependencies

To make good use of the techniques described in the previous section, we need appropriate ***prior distributions*** to compute association with and rely upon in the event of

complete sensor failure or occlusion. Therefore, for the tracking tasks, we take into account temporal context. In addition to object location, the observation association itself may be correlated in time. For example, if the target passes behind an occluder, it may be some time before it becomes visible again on the other side. To model data with these correlations, we introduce the graphical model of Figure 2.4(a), in which the state l and model variables M_i are each now connected through time. To generate from this model, at each time t the location and model variables are selected on the basis of their states at the previous time and the transition probabilities $p(l^{t+1}|l^t)$ and $p(\mathrm{M}_i^{t+1}|\mathrm{M}_i^t)$. Conditional on these variables, each observation is then generated in the same way as for the previous independently and identically distributed (IID) case. Inference may then consist of computing the posterior over the latent variables at each time t given all T available observations, $p(l^t, \mathrm{M}^t | x_1^{1:T}, x_2^{1:T})$ (i.e., smoothing) if processing is off-line. If the processing must be on-line, the posterior over the latent variables given all the data up to the current time $p(l^t, \mathrm{M}^t | x_1^{1:t}, x_2^{1:t})$ (i.e., filtering) may be employed. Multimodal source tracking is performed by computing the posterior of l, marginalizing over possible associations. We have seen previously that the posterior distribution over location at a given time is potentially non-Gaussian (Figure 2.3(d)). To represent such general distributions, one approach is simply to discretize the state space of l. In this case, the dynamic Bayesian network illustrated in Figure 2.3(a) is a factorial hidden Markov model (FHMM) [Ghahramani and Jordan, 1997] configured for data association. See Appendix A.1.1 for details about FHMMs and the inference derivations. In this example, exact numerical inference on the discretized distribution is tractable. Given state transition matrices $p(l^{t+1}|l^t)$ and $p(\mathrm{M}^{t+1}|\mathrm{M}^t)$, we can write down recursions for inference in this FHMM in terms of the posteriors $\alpha^t \triangleq p(l^t, \mathrm{M}_{1,2}^t | D^{1:t})$ and $\gamma^t \triangleq p(l^t, \mathrm{M}_{1,2}^t | D^{1:T})$:

$$\alpha^t \propto \sum_{l^{t-1}, \mathrm{M}_{1,2}^{t-1}} p(D^t | l^t, \mathrm{M}_1^t, \mathrm{M}_2^t) p(l^t | l^{t-1}) \prod_{i=1}^{2} p(\mathrm{M}_i^t | \mathrm{M}_i^{t-1}) \alpha^{t-1}, \tag{2.6}$$

$$\gamma^t \propto \sum_{l^{t+1}, \mathrm{M}_{1,2}^{t+1}} \frac{p(l^{t+1}|l^t) \prod_{i=1}^{2} p(\mathrm{M}_i^t | \mathrm{M}_i^{t-1}) \alpha^t}{\sum_{l^t, \mathrm{M}_{1,2}^t} p(l^{t+1}|l^t) \prod_{i=1}^{2} p(\mathrm{M}_i^t | \mathrm{M}_i^{t-1}) \alpha^t} \gamma^{t+1}. \tag{2.7}$$

Filtering makes use of the forward α recursion in eq. (2.6) and smoothing the backward γ recursion in eq. (2.7), which are analogues of the α and γ recursions in standard hidden Markov model (HMM) inference ([Bishop, 2006a]).

The benefits of temporal context for inference of source state and data association

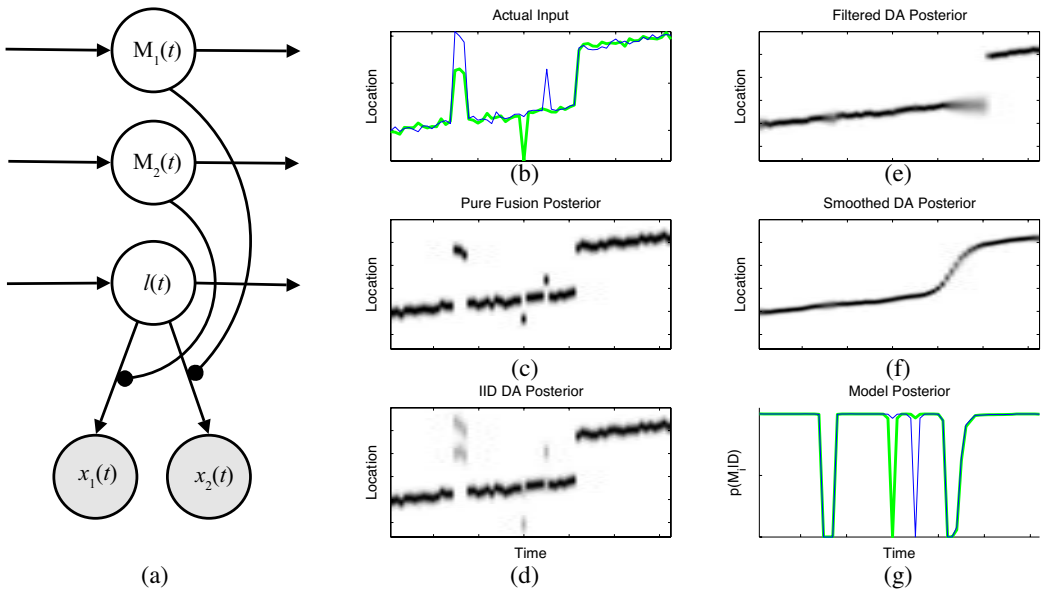


Figure 2.4: (a) Graphical model to describe generation of observations x_i with temporal dependency. (b) Synthetic input data-set in two modalities. Posterior probability of l in (c) pure fusion model (d) IID data association model (e) filtered data association model and (f) smoothed data association model. (g) Posterior probability of model structure for the smoothed data association model.

are illustrated in Figures 2.4(b)-(g). Figure 2.4(b) illustrates data from a series of T observations, $D = \{x_1^t, x_2^t\}_{t=1}^T$, $x_i^t \sim \mathcal{N}(l^t, p)$, in two independent modalities, of a continuously varying latent source l. These data include some occlusions/sensor failures (where the observation(s) are generated from a background distribution) and an unexpected discontinuous jump of the source. The temporal state evolution models for l and M are simple diffusion models.

- [Figure 2.4(c)] A *pure fusion model* without temporal context has very limited robustness, as inference in this model always consists of a simple precision weighted average over observations. This procedure is not useful since the disassociated observations can come from an entirely different distribution and hence, throw off the average entirely.

- [Figure 2.4(d)] A *data association model* is slightly more robust, correctly inferring that the pure fusion generative structure is unlikely when the observations are discrepant. However, without temporal context, it cannot identify which observation was discrepant. Marginalizing over the models, it produces a non-Gaussian, multimodal posterior for l.

- [Figure 2.4(e)] Including some temporal history, an on-line *'filtered' data association model* can infer which observations are discrepant and discount them, producing much smoother inference. In this case, after the discontinuity in state, the fully disassociated observation structure is inferred. Based on the temporal diffusion model, an approximately constant location is inferred until enough evidence is accumulated to support the new location.

- [Figure 2.4(f)] Finally, an off-line *'smoothing' data association model* infers a robust, accurate trajectory. For this case, the marginal posterior of the association variables is shown in Figure 2.4(g).

The illustrative scenarios discussed here generalise in the obvious way to more observations. With many sensors, the disassociation of a small number of discrepant sensors can be inferred even without prior information. However, in a pure fusion scheme, even with many sensors, a single highly discrepant sensor can throw off all the others during averaging.

We have illustrated temporal state inference with unknown data association by discretizing and computing on the entire state space. This is because the main machine learning application (Chapter 3) described in this thesis also maintains a discrete state space representation.

If a continuous state space is necessary, (e.g., because the dimensionality of the data is too large for exhaustive discretization of the state space), then other approaches must be employed. Since the posteriors under uncertain data association are non-Gaussian (Figure 2.3), the classical Kalman filter [Kalman, 1960, Bishop, 2006a] recursions for tracking — which require Gaussian priors and posteriors — are not directly applicable and must be used in conjunction with some approximation . Gaussian sum approximations can be employed to collapse the non-Gaussian posterior into an appropriately corrected single Gaussian at each time-step, e.g., [Bar-Shalom et al., 2005]. To improve on this accuracy with more computational expense, the mixture of Gaussians required to represent observations from a finite window of time-steps can be maintained, e.g., [Williams et al., 2006]. Alternately, sampling based approximations such as particle filtering [Blake and Isard, 1997] can be used for maintaining non-parametric distributions over the state during tracking [Gatica-Perez et al., 2007, Williams et al., 2006].

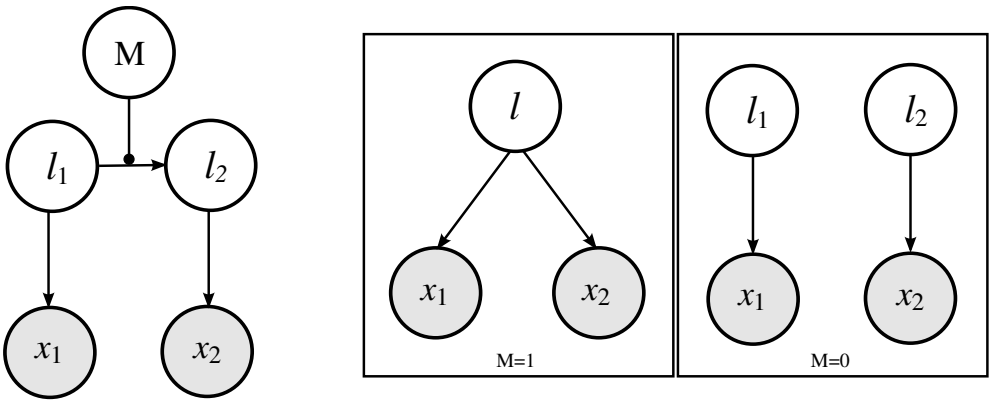


Figure 2.5: Graphical models to describe the generation of multimodal observations x_1, x_2 which may by due to separate sources or one single source. (a) variable structure representation (b) variable model representation.

2.2 Modelling Multiple Sources

There is another simple way in which two multimodal point observations can be generated, i.e., each could be generated by a *separate* source instead of a single source.

2.2.1 An Illustrative Example

The choice of the multi versus single source generating model (Figure 2.5(b)) can also be expressed compactly as structure inference (Figure 2.5(a)) as before, but by using two latent state variables, and requiring equality between them if $\mathrm{M} = 1$ and independence if $\mathrm{M} = 0$. It is possible to enumerate all five possible model structures and perform the Bayesian model selection given the data. However, frequently the semantics of a given perceptual problem correspond to a prior over models which either allows the four discussed earlier ("occlusion semantic") or a choice between one or two sources ("multi-object semantic"). The occlusion semantic arises for example, in audio-visual processing where a source may independently be either visible or audible. The multi-object semantic arises, for example, in some psychophysics experiments [Shams et al., 2000] discussed later.

We will now illustrate the latter case with a toy but concrete example of generating observations in two different modalities x_1, x_2 which may both be due to a single latent source $(\mathrm{M} = 1)$, or two separate sources $(\mathrm{M} = 0)$. Using vector notation, the likelihood of the observation $\mathbf{x} = [x_1, x_2]^T$ given the latent state $\mathbf{l} = [l_1, l_2]^T$ is $\mathcal{N}(\mathbf{x}|\mathbf{l}, \mathbf{P}_x)$ where

$\mathbf{P}_x = diag([p_1, p_2])$. Let us assume the prior distributions over the latent locations are Gaussian but tend toward being informative. In the multi-object model, the prior over l_is: $p(\mathbf{l}|\mathrm{M}=0) = \mathcal{N}(\mathbf{l}|\mathbf{0},\mathbf{P}_0)$ is uncorrelated, so $\mathbf{P}_0 = p_0\mathbf{I}$ and $p_0 \to 0$. In the single object model, the prior over l_is: $p(\mathbf{l}|\mathrm{M}=1) = \mathcal{N}(\mathbf{l}|\mathbf{0},\mathbf{P}_1)$ requires the l_is to be equal, so $\mathbf{P}_1$ is chosen to be strongly correlated. The joint probability of the whole model and the structure posterior are given in eqs. (2.8)-(2.11):

$$
\begin{aligned}
p(\mathbf{x},\mathbf{l},\mathrm{M}) &= \mathcal{N}(\mathbf{x}|\mathbf{l},\mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0},\mathbf{P}_0)^{(1-\mathrm{M})}\mathcal{N}(\mathbf{l}|\mathbf{0},\mathbf{P}_1)^{\mathrm{M}}p(\mathrm{M}), && (2.8)\\
p(\mathrm{M}|\mathbf{x}) &\propto \int_{\mathbf{l}} \mathcal{N}(\mathbf{x}|\mathbf{l},\mathbf{P}_x)\mathcal{N}(\mathbf{l}|\mathbf{0},\mathbf{P}_0)^{(1-\mathrm{M})}\mathcal{N}(\mathbf{l}|\mathbf{0},\mathbf{P}_1)^{\mathrm{M}}p(\mathrm{M}), && (2.9)\\
p(\mathrm{M}=0|\mathbf{x}) &\propto \mathcal{N}(\mathbf{x}|\mathbf{0},(\mathbf{P}_x^{-1}+\mathbf{P}_0^{-1})^{-1})p(\mathrm{M}=0), && (2.10)\\
p(\mathrm{M}=1|\mathbf{x}) &\propto \mathcal{N}(\mathbf{x}|\mathbf{0},(\mathbf{P}_x^{-1}+\mathbf{P}_1^{-1})^{-1})p(\mathrm{M}=1). && (2.11)
\end{aligned}
$$

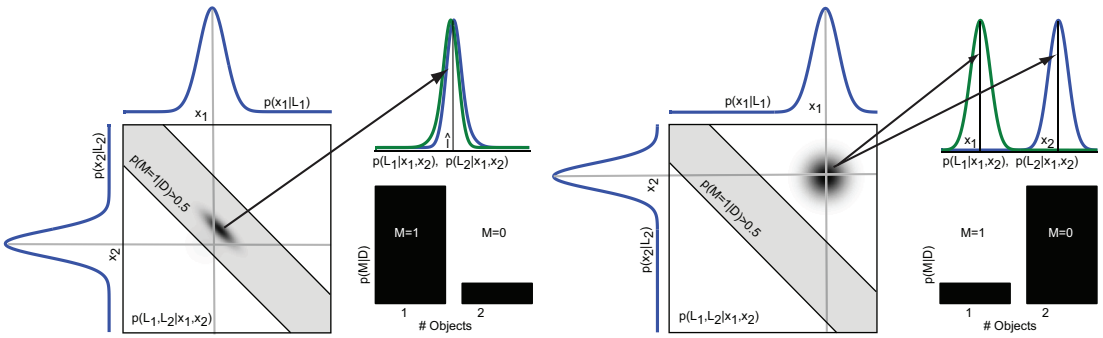
A compact illustration of the interesting behaviours exhibited is show in Figure 2.6.

- [Figure 2.6(a)] If observations x_1 and x_2 (grey cross-hairs) are only slightly discrepant, then the single object model ($\mathrm{M}=1$) is inferred with high probability. The inferred l_is are pulled toward each other, away from their initial observations x_i. (See Figure 2.6(a), the shaded Gaussian posterior is displaced from cross-hair observations toward the concordant-cues axis.) Specifically, the posterior over $\mathbf{l}$ is strongly correlated and Gaussian about the point of the fused interpretation, i.e., $p(\mathbf{l}|\mathbf{x}) \approx \mathcal{N}(\mathbf{l}|\hat{\mathbf{l}},\mathbf{P}_{l|x})$ where $\hat{\mathbf{l}} = \mathbf{P}_{l|x}^{-1}\mathbf{P}_x\mathbf{x}$, $\mathbf{P}_{l|x} = \mathbf{P}_x + \mathbf{P}_1$. The location marginals for each l_i are therefore the same and centred at $\hat{\mathbf{l}}$.

- [Figure 2.6(b)] If observations x_1 and x_2 are highly discrepant, then the two object model is inferred with high probability. The inferred l_is do not interact. (See Figure 2.6(b), the shaded Gaussian posterior is not displaced toward the concordant-cues axis, but is aligned with cross-hair observations.) Specifically, the posterior $p(\mathbf{l}|\mathbf{x})$ is spherical and centred at the observations themselves, rather than a single fused estimate; i.e. $\hat{\mathbf{l}} = \mathbf{P}_{l|x}^{-1}\mathbf{P}_x\mathbf{x} \approx \mathbf{x}$, $\mathbf{P}_{l|x} = \mathbf{P}_x + \mathbf{P}_0$.

2.2.2 Connection to Model Selection

The model selection effect of Bayesian Occam's razor [MacKay, 2003] is clear, particularly in the single versus multi-source case, by considering the entire normalised data distribution for the single and multi-target hypotheses (models) in the two dimensional space (x_1, x_2) (Figure 2.7). Consider that the single-source hypothesis ($\mathrm{M}=1$)



(a) Inferring integrative structure from correlated inputs

(b) Inferring segregative structure from decorrelated inputs

Figure 2.6: Inference in multi-object semantic toy model. (a) For correlated inputs, $x_1 \approx x_2$, the presence of one objects is inferred and its location posterior is the probabilistic fusion of the observations. (b) For very discrepant inputs, $x_1 \neq x_2$, the presence of two objects is inferred and the location posterior for each is at the associated observation.

concentrates its probability mass $p(x_1, x_2|\mathrm{M}=1)$ in a narrow region around $x_1 = x_2$ (Figure 2.7(a)); whereas the multi-source hypothesis ($\mathrm{M}=0$) spreads its probability mass $p(x_1, x_2|\mathrm{M}=0)$ more widely around the space (Figure 2.7(b)).

Inference for M shows that $\mathrm{M}=1$ (one source) is more likely when the observations are similar $x_1 \approx x_2$, and that $\mathrm{M}=0$ (two sources) becomes more likely some with some sufficient discrepancy $|x_1 - x_2|$ (Figure 2.7(c)). Note that computing M based on a maximum likelihood (ML) point estimate of l would result in $\mathrm{M}=0$ – the more complex, two source explanation – being being more probable everywhere. (This is because it would be possible to everywhere explain the data by choosing $l_1 = x_1$ and $l_2 = x_2$.) Alternately, choosing a maximum a posteriori (MAP) estimate for can have the opposite effect: of inferring an inappropriately extensive region for $\mathrm{M}=1$. This is because paying the prior "cost" for two separate l_is to explain the data as $\mathrm{M}=0$ is expensive in this uninformative prior scenario.

This complexity control in Bayesian model inference will turn out to be important to explain various interesting results in human psychophysics: including one described briefly in the next section, as well as experiments considered in detail in Chapter 4.

2.2.3 Related Work

A real, albeit discrete domain in which these multi-object association ideas are relevant are the psychophysics experiments reported in [Shams et al., 2000, Shams et al., 2005]. In these experiments, a variable number (1-4) of approximately

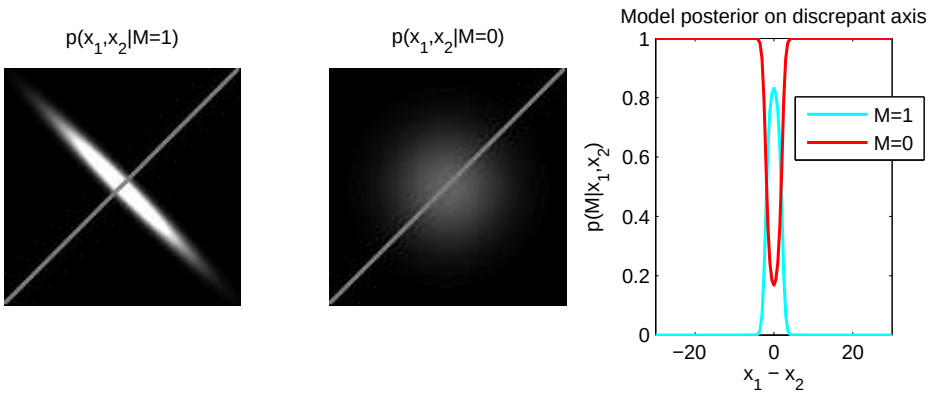


Figure 2.7: Illustration of Bayesian Occam's razor effect in single vs multi source inference in multiple modalities. (a) Single object $(\mathrm{M} = 1)$ data likelihood. (b) Multi object $(\mathrm{M} = 0)$ data likelihood. (c) Model posterior (number of objects) inferred along the discrepant observations diagonal ($x_1 = -x_2$, grey lines).

coincident beeps and flashes are presented to the subject, who must estimate how many were actually presented on the basis of their noisy sensory input. Since in the real world, events frequently produce correlated multimodal observations, the hypothesis that these observations correspond to the same event(s) (Figure 2.5(b), $\mathrm{M} = 1$) is a plausible one for the perceptual system to consider against the hypothesis that they are unrelated (Figure 2.5(b), $\mathrm{M} = 0$).

Based on the model selection described in the previous section, an apparent small discrepancy in the likelihood peaks for beep number and flash number is likely to be due to sensory noise in the observation of a single correlated source (Figure 2.6(a), eq. (2.11)), leading to integration in the perception of the number of beeps and flashes (the perceived number of flashes and beeps tends to be the same). In contrast, an apparent large discrepancy is likely to be because the observations are actually unrelated – caused by two separate sources – (Figure 2.6(b), eq. (2.10)), leading to segregation in the perception of the number of beeps and flashes (the perceived number of flashes and beeps show no interaction). This integration of similar observations, and segregation of highly discrepant observation, is indeed the observed outcome of these experiments. Our interpretation of these experiments is supported by the very recent publication of an analysis ([Kording et al., 2007]) which independently explains these results in the same way, with the same framework presented in this section.

2.3 Discussion

In this chapter, we introduced a structure inference or model selection based theoretical approach for inference of source state and observation association given multisensory observations of uncertain correspondence. This is in contrast to the classical sensor fusion model described in Chapter 1 (Figure 1.1), which assumes that observation association is known. In describing models for inference about single sources (Section 2.1), we showed how our approach can potentially improve on the robustness and accuracy of classical sensor fusion state inference (Figure 2.3). Moreover, we saw that explicit data association can be of use for representing interesting semantic content such as the *number* of sources present (Section 2.2, Figure 2.6). We also discussed how a Bayesian treatment of this structure question can be important for correct inference of association (Section 2.2.2), as illustrated in by the analysis of a recent psychophysics experiment (Section 2.2.3).

All of these properties will be exploited in subsequent chapters. Next, in Chapter 3, we will apply the models for inference about single sources with tracking to real, large scale machine perception problem in the audio-visual domain. Subsequently, in Chapter 4, we will additionally apply the model for inference about multiple sources to understand recent results in human psychophysics.

Chapter 3

Machine Learning of Audio-Visual Scene Understanding

In this chapter, we develop and apply a probabilistic model capable of representing real audio-visual data of variable data association. While being conceptually the same as the toy model in Section 2.1, the audio-visual model is necessarily significantly more detailed than the generic form. In Section 3.1, we introduce the problem setting. Section 3.2 describes the developed model in detail: including algorithms for inference (Section 3.2.2), which will provide detection and tracking; as well as learning (Section 3.2.3), which will provide automatic calibration and user appearance learning. In Section 3.3, we illustrate results for learning, tracking and computing association for human and mechanical targets. We summarize our contributions and their relation to other research in Section 3.4.

3.1 Introduction

To illustrate the application of probabilistic modelling of structure inference to a real, large scale machine perception problem, we consider the task of unsupervised learning and inference with audio-visual (AV) input. Audio-visual scene understanding has most immediate application in teleconferencing [Gatica-Perez et al., 2007, Perez et al., 2004] and machine direction and broadcast [Al-Hames et al., 2006, Zhang et al., 2008] applications. In these scenarios, knowledge of the people's states and locations can be exploited by the system to good effect. For example, by switching cameras, or steering a pan-tilt camera or digital zoom, a system can provide the best view of a speaking lecturer [Zhang et al., 2008] or teleconference

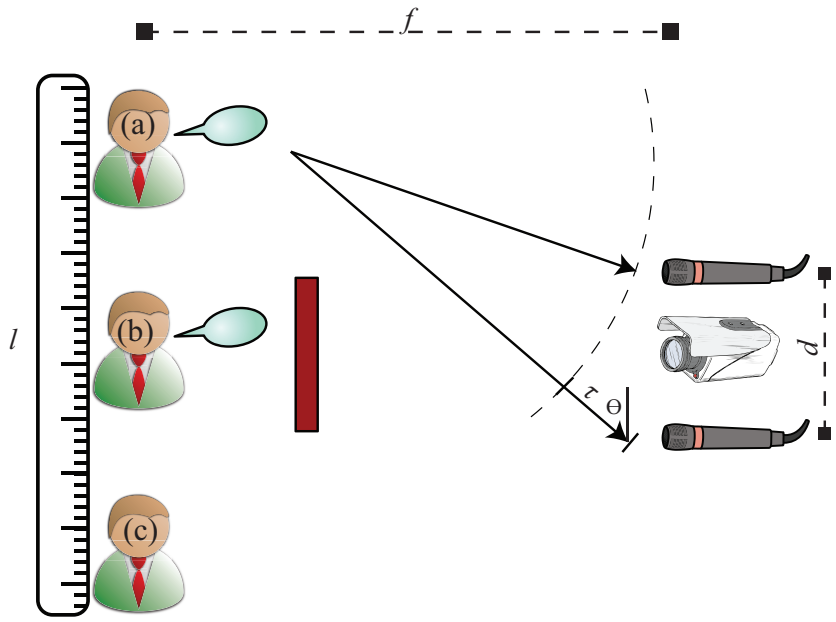


Figure 3.1: Schematic of scenario for audio-visual tracking and scene understanding. User location l is inferred from visual field location when visible, as well as from inter-microphone time delay τ when audible.

participant [Al-Hames et al., 2006] for broadcast. Microphone arrays may also be electronically steered toward a speaker to improve speech reception for broadcast [Zhang et al., 2008]. To be effective, all these applications require accurate knowledge of where people are and their speaking status.

More generally, in AI and cognitive robotics, an overarching goal is to build machine perception systems which can understand and interact the world – requiring general purpose algorithms for learning about objects and inferring their state. For everyday scenarios and interaction with humans, this means *learning* to recognize and locate moving people with vision and audition and, when interacting with multiple humans, it means being able to infer *who said what*.

3.2 Modelling Audio-Visual Scenes

The particular problem scenario we will consider is learning of audio-visual detection, association and tracking (in a fixed azimuthal plane) of single and multiple human users by a system equipped with a digital video camera and microphone pair, as shown schematically by Figure 3.1. Both auditory and visual modalities can potentially provide information about the source location (Figure 3.1(a)).

In the absence of occlusion, correctly detecting known visual features in a given frame of video obviously allows straightforward computation of source location l based on the location of the features within the image. For an audible source, a sequence of samples from a horizontally separated microphone pair can also provide information about the source location. The temporal offset τ between sound wave arrivals at each microphone depends on the azimuthal location of the source. τ can be measured by cross correlation of the signals. It can then be used to compute the angle of incidence θ, and hence the location l of the source. For example, (from the geometry in Figure 3.1) for small microphone separation d, large focal distance f, and speed of sound v; the source location l is approximately given by ([Perez et al., 2004]):

$$l \approx f/\tan\left(\cos^{-1}\left(\frac{\tau v}{d}\right)\right). \tag{3.1}$$

Localization using this cue τ – known as the inter-aural time delay (IATD) – is common in machine perception applications [Perez et al., 2004, Beal et al., 2003], and accounts for a significant fraction of the information used by humans in auditory spatial localization ([Alais and Burr, 2004]).

The AV localization [Perez et al., 2004, Beal et al., 2003] part of the task is similar to the task required in psychophysics experiments such as [Alais and Burr, 2004], where humans are reported to exhibit near Bayes optimal sensor fusion. Existing machine perception work has tended to assume pure fusion and temporal independence (which limits robustness and precludes inferring relational quantities, e.g., [Beal et al., 2003]); require human calibration (which limits usability, e.g., parameters f, d, v in [Perez et al., 2004]); and require human specification of highly constrained recognition models (which limits breadth of application, e.g., facial regions in [Perez et al., 2004]). We now tackle the broader scene understanding problem of *learning* how to *associate* AV data through *time*. By learning the visual appearance, our model will not be constrained to detecting only human faces, and by inferring multisensory data association online, our model should robustly track through occlusion in either modality (Figure 3.1(b),(c)), and ultimately be able to infer *who said what*.

The overall parametric form will be that of the *transformed mixture of Gaussians* (TMG) framework [Frey and Jojic, 2003, Beal et al., 2003], which as we shall see, allows efficient inference and learning with the expectation-maximization (EM) algorithm [Dempster et al., 1977, Bishop, 2006a]. The following sections describe the model, learning, and inference procedures in detail.

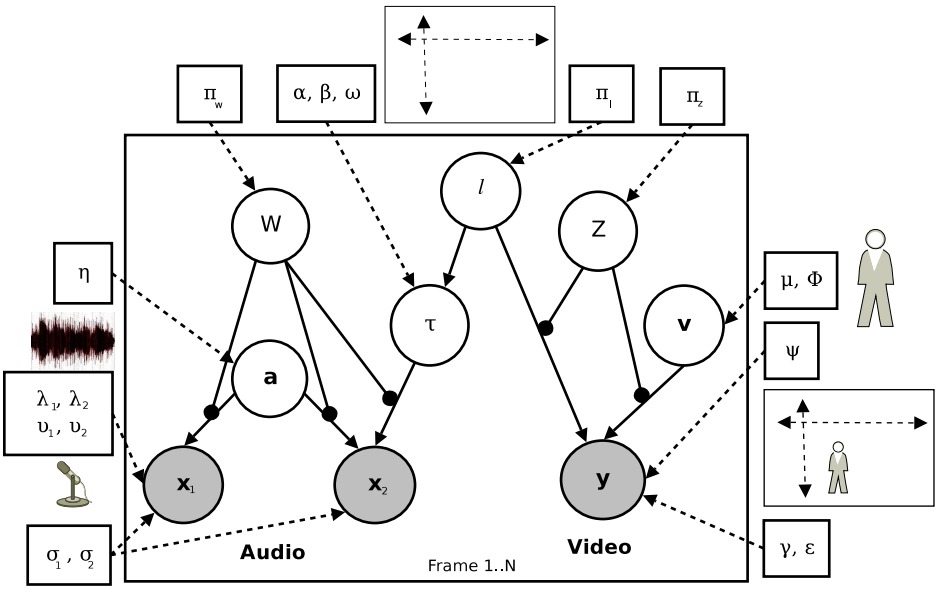


Figure 3.2: Graphical model for audio-visual data generation. Refer to Table 3.1 for summary of notation.

3.2.1 Generative Model

Given the problem scenario illustrated in Figure 3.1, a graphical model to describe the generation of a *single* frame of AV data $D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$[1] is illustrated in Figure 3.2. Table 3.1 summarises the notation used. The generative process can be described as follows: A discrete translation l representing the source state is selected from its prior distribution π_l, and its observability in each modality (W, Z) is selected from its binomial prior. For simplicity, and due to the nature of our data, we only consider source translation along the azimuth in our experiments, so l effectively ranges over all the x-axis pixels of the image[2]. This could easily be expanded to include y-axis translation as in [Beal et al., 2003][3]. First, consider the all visible (pure fusion) case $(\mathrm{W}, \mathrm{Z} = 1)$. The video appearance $\mathbf{v}$ is sampled from a diagonal precision Gaussian distribution $\mathcal{N}(\mathbf{v}|\mu, \Phi)$ with parameters defining its soft template. The observed video pixels are generated by sampling from the spherical Gaussian $\mathcal{N}(\mathbf{y}|\mathbf{T}_l\mathbf{v}, \Psi\mathbf{I})$, the mean of which is the sampled appearance $\mathbf{v}$ translated by l using the transformation matrix $\mathbf{T}_l$. The

[1] $\mathbf{y}$ is a 12,000 element vector representing the raster scanned contents of a given 120x100 pixel frame of video, which is captured at 12.5fps. $\mathbf{x}_i$ are vectors of the 1280 samples recorded at 16kHz from each microphone i during the arrival of each corresponding video frame $\mathbf{y}$.

[2] As l is actually a translation, its range during tracking can be constrained to the region around the current location for computational efficiency[Jojic et al., 2000], however we have not needed to do this.

[3] For tracking vertical and horizontal movement, l would be a two-element vector specifying the horizontal and vertical translations.

Variables		Parameters	
$\mathbf{x}_1, \mathbf{x}_2$	Observed signal at microphones	υ_1, υ_2	Precision of microphone speech reception
		λ_1, λ_2	Microphone speech attenuation
		σ_1, σ_2	Precision of background noise
$\mathbf{a}$	Speech signal	η	Precision of speech signal
τ	Inter microphone time delay	α, β, ω	Determine time delay as linear function of source location
W	Audio association	π_w, Θ	Prior and dynamic probability of audio association
$\mathbf{y}$	Observed video frame	Ψ	Foreground precision of video camera
		γ, ε	Video background mean and precision
$\mathbf{v}$	Video appearance	μ, Φ	Mean and precision of foreground video appearance
Z	Video association	π_z, Ω	Prior and dynamic probability of video association
l	Discrete source location (pixels)	π_l, Γ	Prior and dynamic probability of source location

Table 3.1: Summary of audio-visual model variables and parameters.

latent audio signal $\mathbf{a}$ is sampled from a zero mean, spherical Gaussian, i.e., $\mathcal{N}(\mathbf{a}|\mathbf{0},\eta\mathbf{I})$. (This model can potentially use a Toeplitz matrix η to represent spectral structure in the signal [Beal et al., 2003], but for simplicity we consider the spherical case here.) The time delay τ between the signals at each microphone (eq. (3.1)) is approximated as a linear function of the translation of the source $\mathcal{N}(\tau|\alpha l+\beta,\omega)$. Given the latent signal and the delay, the observation $\mathbf{x}_i$ at each microphone is generated by sampling from a spherical Gaussian with the mean $\mathbf{a}$; with $\mathbf{x}_2$ shifted τ samples relative to $\mathbf{x}_1$, i.e., $\mathbf{x_1} \sim \mathcal{N}(\mathbf{x}_1|\mathbf{a},\upsilon_1\mathbf{I})$, $\mathbf{x_2} \sim \mathcal{N}(\mathbf{x}_2|\mathbf{T}_\tau\mathbf{a},\upsilon_2\mathbf{I})$. If the video modality is occluded ($\mathrm{Z}=0$), the observed video pixels are drawn from a very generic Gaussian background distribution $\mathcal{N}(\mathbf{y}|\gamma\mathbf{1},\varepsilon\mathbf{I})$ independently of l and audio data. If the audio modality is silent ($\mathrm{W}=0$), the samples at each speaker are drawn from very generic background distributions $\mathcal{N}(\mathbf{x}_i|\mathbf{0},\sigma_i\mathbf{I})$ independently of each other, l and the video.

To describe the generation of a series of correlated frames, the IID observation model in Figure 3.2 is replicated at every time-step and a factored Markov model is defined over the location and association variables $(l,\mathrm{W},\mathrm{Z})$ exactly as the toy model was developed previously (refer to Figure 2.4(a)). The state evolution over the location shift is defined in the standard way: $p(l^{t+1}|l^t)=\Gamma_{[l^t,l^{t+1}]}$, where the subscripts pick out the appropriate element of the matrix Γ. The observability transitions are defined similarly as $p(\mathrm{W}^{t+1}|\mathrm{W}^t)=\Theta_{[\mathrm{W}^t,\mathrm{W}^{t+1}]}$ and $p(\mathrm{Z}^{t+1}|\mathrm{Z}^t)=\Omega_{[\mathrm{Z}^t,\mathrm{Z}^{t+1}]}$. Suppressing unambiguous indexing by t for clarity, the joint probability of the model including all visible $D=\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}\}_{t=1}^T$ and hidden variables $H=\{\mathbf{a},\mathbf{v},\tau,l,\mathrm{W},\mathrm{Z}\}_{t=1}^T$ given all the parameters $\theta=\{\lambda_{1,2},\upsilon_{1,2},\eta,\alpha,\beta,\omega,\pi_l,\mu,\Phi,\Psi,\Gamma,\Theta,\Omega,\pi_w,\pi_z,\gamma,\varepsilon,\sigma_{1,2}\}$ factorizes as:

$$
\begin{aligned}
p(D,H|\theta) &= \prod_{t=0}^{T-1} p(l^{t+1}|l^t)p(\mathrm{W}^{t+1}|\mathrm{W}^t)p(\mathrm{Z}^{t+1}|\mathrm{Z}^t) \\
&\quad \cdot \prod_{t=1}^{T} p(\mathbf{x}_1|\mathrm{W},\mathbf{a})p(\mathbf{x}_2|\mathrm{W},\mathbf{a},\tau)p(\mathbf{a})p(\tau|l)p(\mathbf{v})p(\mathbf{y}|\mathrm{Z},\mathbf{v},l), \\
&= \left(\prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_1|\mathbf{a},\upsilon_1)^{w}\mathcal{N}(\mathbf{x}_1|\mathbf{0},\sigma_1)^{\overline{w}}\mathcal{N}(\mathbf{a}|0,\eta) \right. \\
&\quad \cdot \mathcal{N}(\mathbf{x}_2|\mathbf{T}_\tau\mathbf{a},\upsilon_2)^{w}\mathcal{N}(\mathbf{x}_2|\mathbf{0},\sigma_2)^{\overline{w}}\mathcal{N}(\tau|\alpha l+\beta,\omega) \\
&\quad \left. \cdot\, \mathcal{N}(\mathbf{y}|\mathbf{T}_l\mathbf{v},\Psi)^{z}\mathcal{N}(\mathbf{y}|\gamma,\varepsilon)^{\overline{z}}\mathcal{N}(\mathbf{v}|\mu,\Phi) \right) \\
&\quad \cdot \prod_{t=0}^{T-1} \Gamma_{l^t,l^{t+1}}\Theta_{\mathrm{W}^t,\mathrm{W}^{t+1}}\Omega_{\mathrm{Z}^t,\mathrm{Z}^{t+1}} \,. \qquad (3.2)
\end{aligned}
$$

For convenient reference, all of the variables and parameters used in the model are summarized in Table 3.1.

3.2.2 Inference

Let us first consider inference given a single frame of data. The Bayesian network described so far gives us the joint probability in eq. (3.2). Due to the structure of the model, the full posterior over all the latent variables for a single frame of data factors into independently computable terms:

$$p(\mathbf{a},\mathbf{v},\tau,l,\mathrm{W},\mathrm{Z}|\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}) = p(\mathbf{a}|\tau,\mathrm{W},D)p(\mathbf{v}|l,\mathrm{Z},D)p(\tau|l,\mathrm{W},D)p(l,\mathrm{W},\mathrm{Z}|D). \quad (3.3)$$

The quantities of ultimate interest for this audio-visual scene understanding task are the ***location of the source*** and its ***visibility and audibility***. The posterior over these quantities is contained in the factor $p(l,\mathrm{W},\mathrm{Z}|D)$, which is all that need to be computed for efficient performance once the model is trained. However, during the training phase, it will be necessary to compute each component of the full posterior for learning with the EM algorithm. The factors $p(\mathbf{a}|\tau,\mathrm{W},D)$ and $p(\mathbf{v}|l,\mathrm{Z},D)$ are the distributions over the latent signals before noise and will be used to train the audio and video recognition models respectively during the M step. The audio signal posterior could also serve as the input to any other downstream audio processing, for example, speech recognition. Finally, the joint posterior over time delay and location is contained in the product $p(\tau,l,\mathrm{W},\mathrm{Z}|D) = p(\tau|l,\mathrm{W},\mathrm{Z},D)p(l,\mathrm{W},\mathrm{Z}|D)$, which will be used to train the AV link parameters, $\{\alpha,\beta,\omega\}$.

3.2.2.1 Latent Signal Posteriors

In this section, we derive the posteriors over the latent variables which will be necessary for training the models of the audio and video signals as well as the audio-visual link parameters. These are all conditioned on the location l (or delay τ) and observability $z \equiv (Z=1)$ or $w \equiv (W=1)$ in the relevant modality.

The joint distribution over the current video data $\mathbf{y}$ and appearance $\mathbf{v}$ is the product of Gaussians $p(\mathbf{y},\mathbf{v}|l,z) = p(\mathbf{y}|\mathbf{v},l,z)p(\mathbf{v})$ and hence, also Gaussian. Conditioning on the data $\mathbf{y}$, the posterior $p(\mathbf{v}|l,z,\mathbf{y})$ of the current video appearance is Gaussian with statistics $p(\mathbf{v}|l,z,\mathbf{y}) = \mathcal{N}(\mathbf{v}|\mu_{\mathbf{v}|\mathbf{y},l,z},\nu_{\mathbf{v}|z})$, where

$$\mu_{\mathbf{v}|\mathbf{y},l,z} = \nu_{\mathbf{v}|z}^{-1}(\Phi\mu + \mathbf{T}_l^T\Psi\mathbf{y}), \quad (3.4)$$

$$\nu_{\mathbf{v}|z} = \Phi + \Psi. \quad (3.5)$$

$p(\mathbf{v}|l,z,\mathbf{y})$ is the inference about the source's appearance before being corrupted by noise Ψ and translation $\mathbf{T}_l$. (See eqs. (A.5) and (A.6) in Appendix A.1.2 for derivation

and details.) For the purpose of video enhancement, the mean $\mu_{\mathbf{v}|\mathbf{y},l,z}$ of this distribution can be interpreted as the de-noised estimate of the image with foreground obstructions removed [Frey and Jojic, 2003]. It is therefore intuitive that this will later be used to train the video appearance parameters (μ, Φ) during learning (eq. (3.22)).

Similarly to the structure for video, the joint distribution over the current audio data $\mathbf{x}_1, \mathbf{x}_2$ and latent signal $\mathbf{a}$ is the product of Gaussians, $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}|\tau, w) = p(\mathbf{x}_1|\mathbf{a}, w)p(\mathbf{x}_2|\mathbf{a}, \tau, w)p(\mathbf{a})$. Conditioning on the data $\mathbf{x}_1, \mathbf{x}_2$, the posterior $p(\mathbf{a}|\tau, w, \mathbf{x}_1, \mathbf{x}_2)$ is Gaussian with statistics $p(\mathbf{a}|\mathbf{x}, \tau, w) = \mathcal{N}(\mathbf{a}|\mu_{\mathbf{a}|\mathbf{x},\tau,w}, \nu_{\mathbf{a}|w})$, where

$$\mu_{\mathbf{a}|\mathbf{x},\tau,w} = \nu_{\mathbf{a}|w}^{-1}(\lambda_1 \upsilon_1 \mathbf{x}_1 + \lambda_2 \upsilon_2 \mathbf{T}_\tau^T \mathbf{x}_2), \tag{3.6}$$

$$\nu_{\mathbf{a}|w} = \eta + \lambda_1^2 \upsilon_1 + \lambda_2^2 \upsilon_2. \tag{3.7}$$

The mean $\mu_{\mathbf{a}|\mathbf{x},\tau,w}$ represents the best estimate for the true speech signal. (See Appendix A.1.2.3, eqs. (A.12) and (A.13) for derivation and details.)

The posterior $p(\tau|w, l, \mathbf{x}_1, \mathbf{x}_2)$ over the inter-aural time delay τ is a discrete distribution which turns out to be closely related to the cross correlation $\sum_i \mathbf{x}_1[i]\mathbf{x}_2[i+j]$ between the signals. It can be derived in terms of the audio parameters $\lambda_1, \lambda_2, \upsilon_1, \upsilon_2$, the generative delay model $p(\tau|l) = \mathcal{N}(\tau|\alpha l + \beta, \omega)$, and the sufficient statistic $\nu_{\mathbf{a}|w}$ as the following Gaussian product integral[4]:

$$p(\tau|w, l, \mathbf{x}_1, \mathbf{x}_2) = \frac{\int_{\mathbf{a}} p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{a}, \tau|w, l)}{p(\mathbf{x}_1, \mathbf{x}_2|l, w)}, \tag{3.8}$$

$$\propto \int_{\mathbf{a}} p(\mathbf{x}_1|\mathbf{a}, w)p(\mathbf{x}_2|\mathbf{a}, \tau, w)p(\mathbf{a})p(\tau|l), \tag{3.9}$$

$$\log p(\tau|w, l, \mathbf{x}_1\mathbf{x}_2) = \log p(\tau|l) + \log p(\mathbf{x}_1, \mathbf{x}_2|w, l) + K, \tag{3.10}$$

$$= \log p(\tau|l) + \frac{1}{2}\mu_{\mathbf{a}|t,\mathbf{x},w}^T \nu_{\mathbf{a}|w} \mu_{\mathbf{a}|t,\mathbf{x},w} + K, \tag{3.11}$$

$$= \log p(\tau|l) + \lambda_1\lambda_2\upsilon_1\upsilon_2\nu_{\mathbf{a}|w}^{-1} \sum_i \mathbf{x}_1[i]\mathbf{x}_2[i+\tau] + K. \tag{3.12}$$

Since this is a discrete distribution, which can be normalized numerically, we only need to take into account terms dependent on τ.

3.2.2.2 Marginal Observation Likelihoods

The marginal observation likelihoods for each modality, i.e., video $p(\mathbf{y}|\mathrm{Z}, l)$ and audio $p(\mathbf{x}_1, \mathbf{x}_2|\mathrm{W}, l)$, will prove convenient to have at hand when computing the final posterior quantity $p(l, \mathrm{W}, \mathrm{Z}|D)$. We therefore derive them in this section. These likelihoods

[4] See Appendix A.1.2.4, eq. (A.16) for derivation and details.

are useful for thinking about data association. Although they are now functions of a discrete l, these likelihoods are analogous to the Gaussian observation likelihoods introduced in Section 2.1.

For a visible ($z \equiv (Z = 1)$) target, the marginal video likelihood $p(\mathbf{y}|z,l)$ is derived from the jointly Gaussian $p(\mathbf{y},\mathbf{v}|z,l) = p(\mathbf{y}|\mathbf{v},l,z)p(\mathbf{v})$. Integrating out the video appearance $\mathbf{v}$, we have $p(\mathbf{y}|z,l) = \int_{\mathbf{v}} p(\mathbf{y},\mathbf{v}|z,l)$, which is Gaussian with statistics $\mathcal{N}(\mathbf{y}|\mu_{\mathbf{y}|l,z}, \nu_{\mathbf{y}|l,z})$, where

$$\mu_{\mathbf{y}|l,z} = \mathbf{T}_l \mu, \tag{3.13}$$

$$\nu_{\mathbf{y}|l,z} = (\Psi^{-1} + \mathbf{T}_l \Phi^{-1} \mathbf{T}_l^T)^{-1}. \tag{3.14}$$

(See eqs. (A.8) and (A.9) in Appendix A.1.2.2 for derivation and details.)

Video disassociation $\bar{z}$ could be due to various causes, including absence of the target, occlusion by another object, or sensor failure. The likelihood of the data given $\bar{z}$ is therefore defined by a very general background distribution:

$$p(\mathbf{y}|l,\bar{z}) = \mathcal{N}(\mathbf{y}|\gamma\mathbf{1}, \varepsilon\mathbf{I}). \tag{3.15}$$

Note that this is now independent of location l. For the background video distribution, a more structured, diagonal Gaussian, precision matrix is also possible, but the more generic spherical Gaussian will turn out to be more useful in Section 3.3.3.

For an audible $w \equiv (\mathrm{W} = 1)$ target, the marginal likelihood $p(\mathbf{x}_1,\mathbf{x}_2|\tau,w)$ is also derived from the jointly Gaussian $p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{a}|\tau,w) = p(\mathbf{x}_1|\mathbf{a},w)p(\mathbf{x}_2|\mathbf{a},\tau,w)p(\mathbf{a})$. Integrating out the audio signal $\mathbf{a}$, we have $p(\mathbf{x}_1,\mathbf{x}_2|\tau,w) = \int_{\mathbf{a}} p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{a}|\tau,w)$, which is given by

$$p(\mathbf{x}_1,\mathbf{x}_2|\tau,w) = \sqrt{\frac{|\upsilon_1|\,|\upsilon_2|\,|\eta|}{(2\pi)^2\,|v_{\mathbf{a}|w}|}} \exp -\frac{1}{2}\left(\mathbf{x}_1^T v_1 \mathbf{x}_1 + \mathbf{x}_2^T v_2 \mathbf{x}_2 - \mu_{\mathbf{a}|t,\mathbf{x},w}^T \nu_{\mathbf{a}|w} \mu_{\mathbf{a}|t,\mathbf{x},w}\right), \tag{3.16}$$

(See Appendix A.1.2.4, eq. (A.16) for derivation and details.) We are however, ultimately interested in the marginal likelihood given the *location*, $p(\mathbf{x}_1,\mathbf{x}_2|l,w)$. To obtain this from eq. (3.16), we combine it with the posterior over the discrete τ (as computed in eq. (3.12)), and numerically integrate τ out (see eq. (3.17)). Similarly to the video model, the marginal likelihood for background noise - conditioned on audio disassociation $\overline{w}$ - is a simple background distribution independent of l (see eq. (3.18)).

$$p(\mathbf{x}_1,\mathbf{x}_2|l,w) = \sum_{\tau} p(\mathbf{x}_1,\mathbf{x}_2|\tau,w)p(\tau|w,l,\mathbf{x}_1,\mathbf{x}_2), \tag{3.17}$$

$$p(\mathbf{x}_1, \mathbf{x}_2 | l, \overline{w}) \quad = \quad \mathcal{N}(\mathbf{x}_1 | \mathbf{0}, \sigma_1 \mathbf{I}) \mathcal{N}(\mathbf{x}_2 | \mathbf{0}, \sigma_2 \mathbf{I}). \tag{3.18}$$

Note that the background audio likelihood has eliminated the *intra*-modality correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$ (as they are no longer related via $\mathbf{a}$). In an alternate formulation of the audio background distribution, conditioning on disassociation $\overline{w}$ could simply eliminate the *inter*-modality correlation (i.e. by making $p(\tau | l, \overline{w})$ uniform instead of peaked) and index a new precision $\eta_{\overline{w}}$ instead of eliminating the intra-microphone correlation entirely. We will make use of this in Section 3.3.3.

3.2.2.3 Location and association posterior

We can now relate detection and tracking in the more complex AV probabilistic model to the generic cases discussed in Section 2.1. For a single frame, the quantity of interest for this task is that of audibility, visibility and location given the data $p(\mathrm{W}, \mathrm{Z}, l | D)$. This is analogous to the posterior over model and location $p(\mathrm{M}_1, \mathrm{M}_2, l | D)$ discussed in the generic case, where we saw in eq. (2.1) that

$$\begin{aligned} p(\mathrm{M}_1, \mathrm{M}_2, l | D) \quad \propto \quad & \mathcal{N}(x_1 | l, p_1)^{\mathrm{M}_1} \mathcal{N}(x_1 | 0, p_b)^{(1-\mathrm{M}_1)} \mathcal{N}(x_2 | l, p_2)^{\mathrm{M}_2} \mathcal{N}(x_2 | 0, p_b)^{(1-\mathrm{M}_2)} \\ & \cdot \mathcal{N}(l | 0, p_l) p(\mathrm{M}_1) p(\mathrm{M}_2). \end{aligned}$$

With the AV marginal likelihoods eqs. (3.13)-(3.18), as computed in Section 3.2.2.2, we can also compute $p(\mathrm{W}, \mathrm{Z}, l | D)$ analogously as:

$$p(\mathrm{W}, \mathrm{Z}, l | D) \quad \propto \quad p(\mathbf{y} | \mathrm{Z}, l) p(\mathbf{x}_1, \mathbf{x}_2 | l, \mathrm{W}) p(\mathrm{Z}) p(\mathrm{W}) p(l), \tag{3.19}$$

$$= \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y}|l,z}, \nu_{\mathbf{y}|l,z})^{z} \mathcal{N}(\mathbf{y} | \gamma \mathbf{1}, \varepsilon \mathbf{I})^{\overline{z}} \left(\sum_t p(\mathbf{x}_1, \mathbf{x}_2 | \tau, w) p(\tau | w, l) \right)^{w} \tag{3.20}$$

$$\cdot \left(\mathcal{N}(\mathbf{x}_1 | \mathbf{0}, \sigma_1 \mathbf{I}) \mathcal{N}(\mathbf{x}_2 | \mathbf{0}, \sigma_2 \mathbf{I}) \right)^{\overline{w}} p(Z) p(W) p(l) \tag{3.21}$$

When computing the filtered or smoothed posterior from multiple frames in the toy model, we saw that the individual observations could be used with the FHMM recursions eqs. (2.6) and (2.7). In the AV case, the filtered or smoothed posterior $p(\mathrm{W}^t, \mathrm{Z}^t, l^t | D^{1:T})$ is computed analogously by using new marginal likelihoods $p(\mathbf{y} | l, \mathrm{Z})$ and $p(\mathbf{x}_1, \mathbf{x}_2 | l, \mathrm{W})$ in recursions eqs. (2.6) and (2.7). (See Appendix A.1.1 for FHMM recursion details.)

3.2.3 Learning

All the parameters in this model $\theta = \{\lambda_{1,2}, \upsilon_{1,2}, \eta, \alpha, \beta, \omega, \pi_l, \mu, \Phi, \Psi, \Gamma, \Theta, \Omega, \pi_w, \pi_z, \gamma, \varepsilon, \sigma_{1,2}\}$ are jointly optimized by a standard EM procedure [Dempster et al., 1977, Frey and Jojic, 2005]. Inference of the posterior distribution $p(H|D)$ over hidden variables H given the observed data D, (as computed in (3.3)), is alternated with optimization of the expected complete log likelihood with respect to the parameters: $\frac{\partial}{\partial\theta}\int_H p(H|D,\theta)\log p(H,D|\theta)$.

The update for the mean μ of the source visual appearance distribution is given by

$$\mu \leftarrow \sum_{t,l} p(l^t, z^t|D^{1:T})\mu^t_{\mathbf{v}|\mathbf{y},l,z} / \sum_t p(z^t|D^{1:T}). \tag{3.22}$$

This is defined in terms of the posterior mean $\mu^t_{\mathbf{v}|\mathbf{y},l,z}$ of the video appearance given the data D for each frame t and translation l, as inferred during the E step in eq. (3.4). (See eq. (A.18) and Appendix A.1.3 for details and derivation.) Intuitively, the result is a weighted sum of the appearance inferences over all frames and transformations, where the weighting is the posterior probability of transformation and visibility in each frame.

The scalar precision parameter σ_i of the background noise is given by

$$\sigma_i^{-1} \leftarrow \sum_t p(\overline{w}^t|D^{1:T})(\mathbf{x}^t_i)^T\mathbf{x}^t_i / N_f \sum_t p(\overline{w}^t|D^{1:T}), \tag{3.23}$$

where N_f specifies the number of samples per audio frame. (See eq. (A.24) and Appendix A.1.3 for details and derivation.) Again, it is intuitive that the estimate of the background variance should be a weighted sum of square of signals at each frame, where the weighting is the posterior probability that the source was silent in that frame. In an IID context, the posterior over the relevant variables l^t and W^t is only dependent on the current observation D^t; so the marginals $p(w^t|D^t)$ etc. would replace those used above for weighting. A full list of updates is given in Appendix A.1.3.

3.2.4 Computational and Implementation Details

In the following sections we detail how to improve the efficiency of the computationally expensive steps in inference and learning (along the lines of [Frey and Jojic, 2003]), and how to ensure numerical stability during EM convergence.

3.2.4.1 Efficiency

The major computationally intensive steps in the inference are the computation of the observation likelihoods for *every single discrete position* l (eq. (3.13)) or *time delay* τ (eq. (3.16)) and the computation of the posterior over τ (eq. (3.12)). Upon closer inspection, these equations can be re-expressed in terms of correlations and convolutions. This allows efficient computation by fast Fourier transform (FFT) by exploiting the property that cross-correlation of two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is equivalent to simple element-wise multiplication in Fourier domain, $\mathcal{F}[\mathrm{Corr}(\mathbf{x}_1,\mathbf{x}_2)] = \mathcal{F}[\mathbf{x}_1]^* * \mathcal{F}[\mathbf{x}_2]$ (where superscript * indicates the complex conjugate). This replaces the $O(N^2)$ correlation with $O(N \log N)$ FFTs.

For example, consider the posterior over τ. From eq. (3.12), we have:

$$\log p(\tau|w,l,\mathbf{x}_1,\mathbf{x}_2) = \log p(\tau|l) + \lambda_1\lambda_2\upsilon_1\upsilon_2\nu^{-1}_{\mathbf{a}|w}\sum_i \mathbf{x}_1[i]\mathbf{x}_2[i+\tau] + \log K, \quad (3.24)$$

$$= \log p(\tau|l) + \lambda_1\lambda_2\upsilon_1\upsilon_2\nu^{-1}_{\mathbf{a}|w}\mathrm{Corr}[\mathbf{x}_1,\mathbf{x}_2] + \log K. \quad (3.25)$$

Considering the audio likelihood $p(\mathbf{x}_1,\mathbf{x}_2|\tau,w)$, we have from eq. (3.16):

$$\log p(\mathbf{x}_1,\mathbf{x}_2|\tau,w) = \frac{1}{2}\mu^T_{\mathbf{a}|t,\mathbf{x},w}\nu_{\mathbf{a}|w}\mu_{\mathbf{a}|t,\mathbf{x},w} + K,$$

$$= \frac{1}{2}v^{-1}_{\mathbf{a}|w}\sum_i \left(\lambda_1^2\upsilon_1^2\mathbf{x}_1[i]^2 + 2\lambda_1\lambda_2\upsilon_1\upsilon_2\mathbf{x}_1[i]\mathbf{x}_2[i+\tau]\lambda_2^2\upsilon_2^2\mathbf{x}_2[i]^2\right) + K, \quad (3.26)$$

$$= \frac{1}{2}v^{-1}_{\mathbf{a}|w}\sum_i \left(\lambda_1^2\upsilon_1^2\mathbf{x}_1[i]^2 + 2\lambda_1\lambda_2\upsilon_1\upsilon_2\mathrm{Corr}[\mathbf{x}_1,\mathbf{x}_2] + \lambda_2^2\upsilon_2^2\mathbf{x}_2[i]^2\right) + K. \quad (3.27)$$

The expensive quadratic terms involving both sample index i and and delay τ in eqs. (3.24) and (3.26) have been expressed as an efficiently computable correlation in eqs. (3.25) and (3.27).

The learning procedure also involves many potentially computationally expensive steps. For example, updating the video appearance μ in eq. (3.22) requires saving the means of the inferred appearances $\mu_{\mathbf{v}|l,\mathbf{y},z}$ for every possible discrete translation l, and then computing their weighted sum. This potentially requires storage and computation of $O(N^2)$ in the number of pixels N. To re-express this update in terms of convolutions (and hence FFTs) we substitute the video appearance inference statistics of eqs. (3.4) and (3.5) into to the numerator of update eq. (3.22):

$$\sum_{t,l} p(l^t,z^t|D^{1:T})\mu^t_{\mathbf{v}|\mathbf{y},l,z} = \sum_{t,l} p(l^t,z^t|D^{1:T})(\Phi+\Psi)^{-1}(\Phi\mu + \mathbf{T}_l^T\Psi\mathbf{y}^t),$$

$$= (\Phi+\Psi)^{-1}\sum_t\left(p(z^t|D^{1:T})\Phi\mu+\Psi\sum_l p(l^t,z^t|D^{1:T})\mathbf{T}_l^T\mathbf{y}^t\right),$$

$$\left(\sum_{t,l}p(l^t,z^t|D^{1:T})\mu^t_{\mathbf{v}|\mathbf{y},l,z}\right)[i] = (\phi[i]+\psi[i])^{-1} \cdot\sum_t\left(p(z^t|D^{1:T})\phi[i]\mu[i]+\psi[i]\sum_l p(l^t,z^t|D^{1:T})[l]\mathbf{y}^t[i-l]\right), \quad (3.28)$$

where ϕ and ψ index the diagonal elements of Φ and Ψ respectively. The final update is therefore given by:

$$\mu \leftarrow \frac{(\Phi+\Psi)^{-1}\sum_t\left(p(z^t|D^{1:T})\Phi\mu+\psi\text{Conv}\left[p(l^t,z^t|D^{1:T}),\mathbf{y}^t\right]\right)}{\sum_t p(z^t|D^{1:T})}. \quad (3.29)$$

The expensive quadratic term in eq. (3.28) involving both pixel index i and image translation l has been replaced with an efficient $O(N\log N)$ convolution in eq. (3.29). Moreover, re-expressing the update directly in terms of the convolutions of the data rather than inference output $\mu_{\mathbf{v}|l,\mathbf{y},z}$, has eliminated the need to store $\mu_{\mathbf{v}|l,\mathbf{y},z}$, enabling space-efficient learning. The updates for the sensor precision ψ, and template precision ϕ, are expressed as FFTs similarly. (See Appendix A.1.4 for list of all FFT computations.)

3.2.4.2 Numerical Stability

There is one major numerical issue in the algorithm as described so far. We can, for example, compute the log-likelihoods $\log p(\mathbf{y}|l,\mathrm{Z})$ for individual values of Z. However, during early cycles of learning, before the parameters are well refined, the likelihood of one model may be much greater than the other, such that the likelihood $p(\mathbf{y}|l,\mathrm{Z})$, and hence the posterior $p(\mathrm{Z},l|\mathbf{y}) \propto p(\mathbf{y}|l,\mathrm{Z})p(l)p(\mathrm{Z})$, are in danger of underflow for one or other value of Z. Constraining entries in the table $p(\mathrm{Z},l|\mathbf{y})$ to be above a minimum small value during normalization is insufficient: for example, information about the *shape* of the associated log likelihood $\log p(\mathbf{y}|l,z)$ as a function of l would still potentially be lost if $p(\mathbf{y}|l,z) \ll p(\mathbf{y}|\bar{z})$. This shape information is important for updates such as eq. (3.22), which are necessary to properly refine the templates.

To help EM converge in a numerically stable way, for the first few cycles, we therefore modify the computation of log-likelihoods in the E-step to constrain the less likely model to be at most K less likely than the other. That is, if for example, $\log p(\mathbf{y}|l,\bar{z}) > K + \log p(\mathbf{y}|l,z)$, then $\log p(\mathbf{y}|l,z)$ is replaced with $\log p(\mathbf{y}|l,\bar{z}) -$

$\text{argmax}_l\{\log p(\mathbf{y}|l,z)\} - K + \log p(\mathbf{y}|l,z)$. i.e., $p(\mathbf{y}|l,z)$ is not allowed to be more than $\exp(K)$ less likely than $p(\mathbf{y}|l,\bar{z})$. Values of about $K = 10$ seem to be suitable. The same procedure is performed for the audio likelihoods $p(\mathbf{x}_1, \mathbf{x}_2|\tau, \mathrm{W})$.

3.3 Robust Audio-Visual Scene Understanding

In this section, we will present results for unsupervised learning and inference in the model presented in Section 3.2 using real world raw AV data. Inference of the posterior $p(l^t, \mathrm{W}^t, \mathrm{Z}^t|D)$ corresponds to source detection via W and Z, source tracking via l and AV source verification if $w \wedge z$. Unsupervised learning of the video parameters (μ, ϕ) corresponds to learning a soft visual template for the object to be tracked. This is in contrast to many other tracking techniques, which require operator specification of the object to be tracked. Moreover, many other audio-visual multimodal systems require careful calibration of the microphone and camera parameters. In this model, these parameters are encompassed by the model AV link parameters (α, β, ω), which are also learned, rendering the model self calibrating.

3.3.1 Inferring the Behaviour of an AV Source: Detailed Example

Results for an illustrative AV sequence after 25 cycles of EM are illustrated in Figure 3.3. In this sequence, the user is initially walking and talking, is then occluded behind another person while continuing to speak, and then continues to walk while remaining silent. Figure 3.3(a) illustrates three representative video frames from each of these segments with the inferred data association and location superimposed on each. In Figure 3.3(c)-(f), the performance of different variants of the tracking algorithm on this data set are compared. Likelihood and posterior modes rather than full location distributions are shown for clarity. Audio and video likelihood peaks are indicated by circles and triangles respectively. The intrinsic imprecision in the audio likelihood compared to that of the video is clear in their relative spread. In each case, the mode of the final location posterior is indicated by the continuous line.

3.3.1.1 Tracking with IID pure fusion model

In the simplest IID pure fusion model, we constrain $\mathrm{W} = \mathrm{Z} = 1$ and use the prior π_l instead of transition matrix Γ. Notice that this now corresponds to the model of Beal

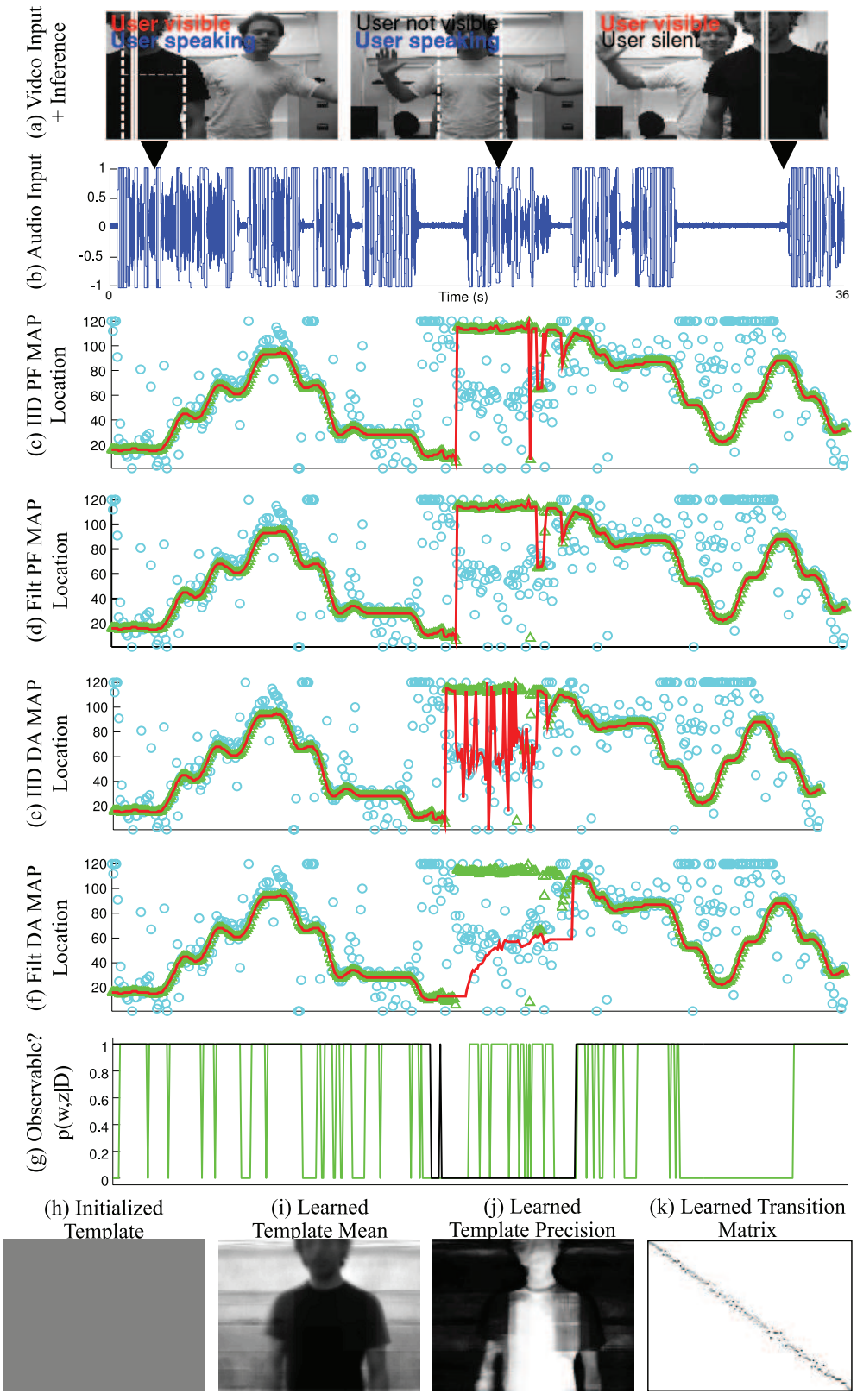


Figure 3.3: AV learning & inference results. (a) Video and (b) audio data with intermittent walking, speaking and occlusion. MAP location with (c) IID pure fusion, (d) filtered pure fusion, (e) IID data association and (f) filtered data association. Likelihood peaks for audio (circles) and for video (triangles). Final output (dark/red line). (g) Visibility (black) and audibility (light/green) posterior. (h) Initial and (i,j) learned video appearance. (k) Learned location transition matrix.

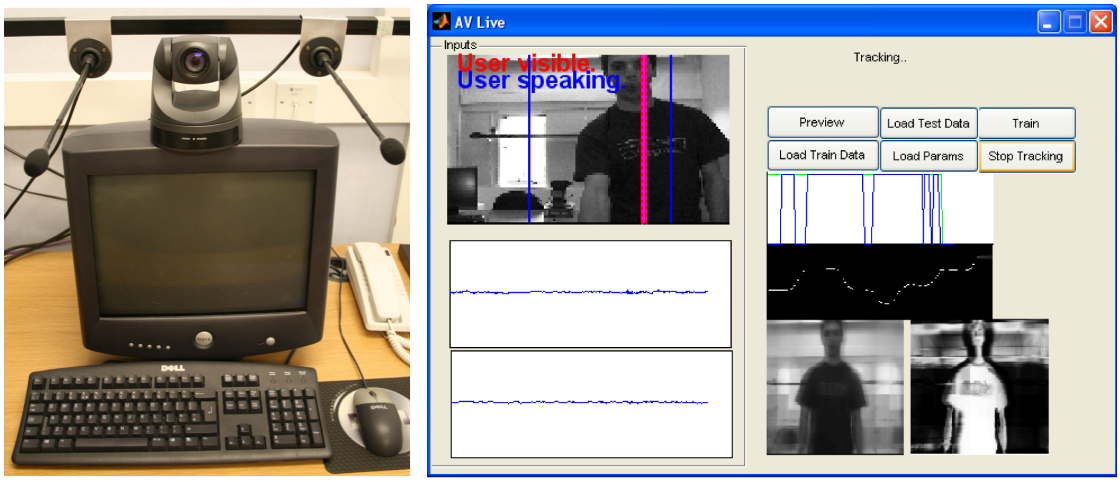
et al. [Beal et al., 2003]. The location inference is correct where the multimodal observations are indeed associated (Figure 3.3(c)). The video modality dominates the fusion as it is much higher precision (i.e., the likelihood function is much sharper), and the posterior is still therefore correct during the visible but silent period where the weaker peaks in the audio likelihood are spurious. While the person in the video foreground is occluded but speaking, the audio likelihood peaks are generally appropriately clustered. However, the next best match to the learnt dark foreground template usually happens to be the filing cabinet in one corner or monitor in the other. With pure fusion, the incorrect but still relatively precise video likelihood dominates the less precise audio likelihood, resulting in a wildly inappropriate posterior.

3.3.1.2 Tracking with filtered pure fusion model

In this case also, we constrain $\mathrm{W},\mathrm{Z} = 1$, but now enable temporal tracking with the transition matrix Γ. This is analogous to the multi-observation Kalman filter – a standard technique for multimodal tracking ([Kalman, 1960, Bishop, 2006a]). Here, a similar type of error as described in the IID pure fusion case is made when filtered tracking is used (Figure 3.3(d)). The only difference is that because of the tracking functionality, the jump between the two incorrect locations is eliminated and the more common of the two previous erroneous locations is focused on.

3.3.1.3 Tracking with IID data association model

In the IID data association model (Figure 3.3(e)), we do not consider temporal tracking, but we do infer W and Z and marginalize over them for localization. The video modality is correctly inferred with high confidence to be disassociated during the occluded period because the template match is poor. The final posterior during this period is therefore based mostly on the audio likelihood, and is generally peaked around the correct central region of the azimuth. The outlier points here have two causes. As speech data is intrinsically intermittent, *both* modalities occasionally have low probability of association, during which times the final estimate is still inappropriately attracted to that of the video modality as in the pure fusion case. Others are simply due to the lower inherent precision of the audio modality.



(a) (b)

Figure 3.4: (a) Computer equipped with camera and microphone pair. (b) User interface for unsupervised appearance learning, audio-visual data association and tracking.

3.3.1.4 Tracking with filtered data association model

In the full data association tracking model, we compute the full posterior $p(l^t, \mathrm{W}^t, Z^t | D^{1:t})$ at every time t. The data association posterior $p(\mathrm{W}^t, Z^t | D^{1:t})$ (Figure 3.3(g)) correctly represents the visibility and audibility of the target at the appropriate times and the information from each of the sensor(s) is appropriately weighted for localization. With the addition of temporal context, tracking based on the noisy and intermittent audio modality is much more reliable in the difficult period of visual occlusion. The user is now reliably and seamlessly tracked during all three domains of the input sequence (Figure 3.3(f)). The inferred data association (Figure 3.3(g)) is used to label the frames in (Figure 3.3(a)) with the user's speaking/visibility status. To cope with intermittent cues, previous multimodal machine perception systems in this context have relied on observations of discrepant modalities providing uninformative likelihoods [Perez et al., 2004, Beal et al., 2003]. This may not always be the case (cf. Figure 3.3(d)), as is evident from our example video sequence where only the data association models succeed during the video occlusion.

Using 120x100 pixel video frames and 1000 sample audio frames, our matlab implementation can perform on-line real time (filtered) tracking at 50fps after learning, which proceeds at 10fps. To use our system, the user approaches an audio-visually equipped PC (Figure 3.4(a)) and presses the train button on our application interface (Figure 3.4(b)), after which he/she is requested to intermittently move around

and speak while 20 seconds worth of training data is collected. After data collection, the EM algorithm is initiated and training takes about five minutes. Once trained, the user is subsequently audio-visually detected, verified and tracked in real time. If the same person is to re-use the system, the parameters of the Bayesian network can be saved and re-loaded later to avoid re-training.

3.3.2 Inferring the Behaviour of an AV Source: Quantitative Evaluation

In the previous section, we described in detail the processing of an example sequence which illustrated most of the important qualitative differences in the behaviour between the model variants. In this section, we describe the results of a more extensive quantitative evaluation of the models against ground truth for a variety of sequences. Rather than using the typical manual markup of video sequence ground truth, we chose to apply the promising but under-explored approach of mechanical generation of test data.

3.3.2.1 Evaluation Procedure

We constructed a computer positionable audio-visual source using an off the shelf speaker component driven on a rail by stepper motor (Figure 3.5). This allowed us to control precisely and repeatably the AV source location along 2.2m of the horizontal plane and to control its audibility and visibility as required for evaluation of audio-visual tracking. This automatic generation of training data provided us with a source of ground truth information without the need for manual labelling. Different visual appearances could be selected by attaching different objects to the movable AV source carriage.

To evaluate the models' performance in variety of different conditions, we controlled four separate variables for a total of 24 different conditions as follows: When present, the audio signal was played at either high or low volume and was composed of low pass filtered noise below either 16,384Hz, 2,048Hz or 256Hz (making audio localization increasingly imprecise). Although somewhat less realistic, a simple noise signal was used rather than recorded speech, so that ground truth audibility could be clearly controlled without the uncertainty in labelling of inter-word pauses etc. [Siracusa and Fisher, 2007]. Indeed, this has resulted in the audio-only tracking output having significantly less variance than, for example, the speech based example illus-

Figure 3.5: Sound positioning device. Position is controllable across 2.2m in the horizontal plane to 1mm accuracy. Inset: The speaker generating the audio source.

trated in Figure 3.3. The visual appearance of a person was simulated by attaching two possible different sets of clothing, and the room lights were either on or dimmed.

The camera's field of view was set up to include the central ~1.5m of the possible source locations. The source speed was up to 0.1ms^{-1} (or ~0.7 pixels per frame in this camera configuration), which produced movement sequences which were slightly easier to track than the human sequences in Sections 3.3.1 and 3.3.3, which tended to have larger velocities and abrupt accelerations.

For each condition, ~40s of training data was collected using three constant velocity passes of the AV source across the field of view of the camera. 25s of same condition test data was then collected using a fixed pattern of movement (see Figure 3.6) including fixed periods of (in)audibility and (in)visibility behind an occluding curtain. We tested performance using two approaches: 1. Same condition testing, where the training data for the matching condition was used to train the model before testing on data from the same condition. 2. Cross condition testing, where the training data for each visual appearance in the easiest condition (lights on, high volume, high frequency) was used to train the model before testing on all the other conditions. These two evaluations quantify two aspects of performance: 1. The models' performance when trained appropriately under actual usage conditions (since a feature of our approach is rapid unsupervised learning of particular contexts). 2. The models' performance when there is deviation between the training and usage scenarios.

3.3.2.2 Evaluation results

We assume for the purposes of evaluation that the probabilistic tracker is required to make a single best guess of every quantity at every time, and take the mode of the posterior distribution output at any point as its best answer. Figure 3.6 details the distribution over the tracker outputs across the 24 test cases with Figure 3.6(a),(b) and Figure 3.6(c),(d) reporting same condition and cross condition testing, respectively. The ground truth position is illustrated by the plain black lines and the ground truth periods of video and audio occlusion are illustrated by the shaded bars below the plots. The distribution of outputs of the audio and video trackers is shown by the light/blue and medium/green shaded regions respectively, and should be interpreted in the context of the occlusion periods. In Figure 3.6(a),(c), the dark/red shaded region illustrates the output distribution for the IID pure fusion model [Beal et al., 2003]. It almost entirely overlaps the (medium/green shaded) video region as the video modality is dominant and deviates drastically from the ground truth (black line) during the entire video occlusion. In Figure 3.6(b),(d) the dark/red shaded region illustrates the output for the filtered data association model developed in this paper. It relatively successfully follows the target using audio only during the initial part of the video occlusion but then fails once the audio occlusion begins. This is because, based on its simple diffusion model of motion, it keeps predicting the same increasingly incorrect location, albeit with decreasing confidence. This continues until the video becomes available again and tracking is regained.

To contrast with the distribution of tracker outputs over trials (Figure 3.6), we also present the full posterior distribution computed by the tracker for a typical same condition trial in Figure 3.7. Inference based on the video modality only (Figure 3.7(a)) focuses on the next best location with confidence during occlusion (see discrepancy between sharp posterior and ground truth in Figure 3.7(a)). (This is the same problem observed in the human trial as discussed in Section 3.3.1.1). In contrast, inference based on the audio modality only (Figure 3.7(b)) is suitably uncertain during the period of audio silence. Combining these two modalities, pure fusion inference (Figure 3.7(c)) is therefore dominated by the video modality and hence is confidently incorrect during the video occlusion. Finally, by inferring also the structure posterior (Figure 3.7(d)), the video modality can be discounted during the occlusion period, and tracking continues based on the less precise audio data only.

The data illustrated in Figure 3.6 are quantified in Table 3.2. For each model vari-

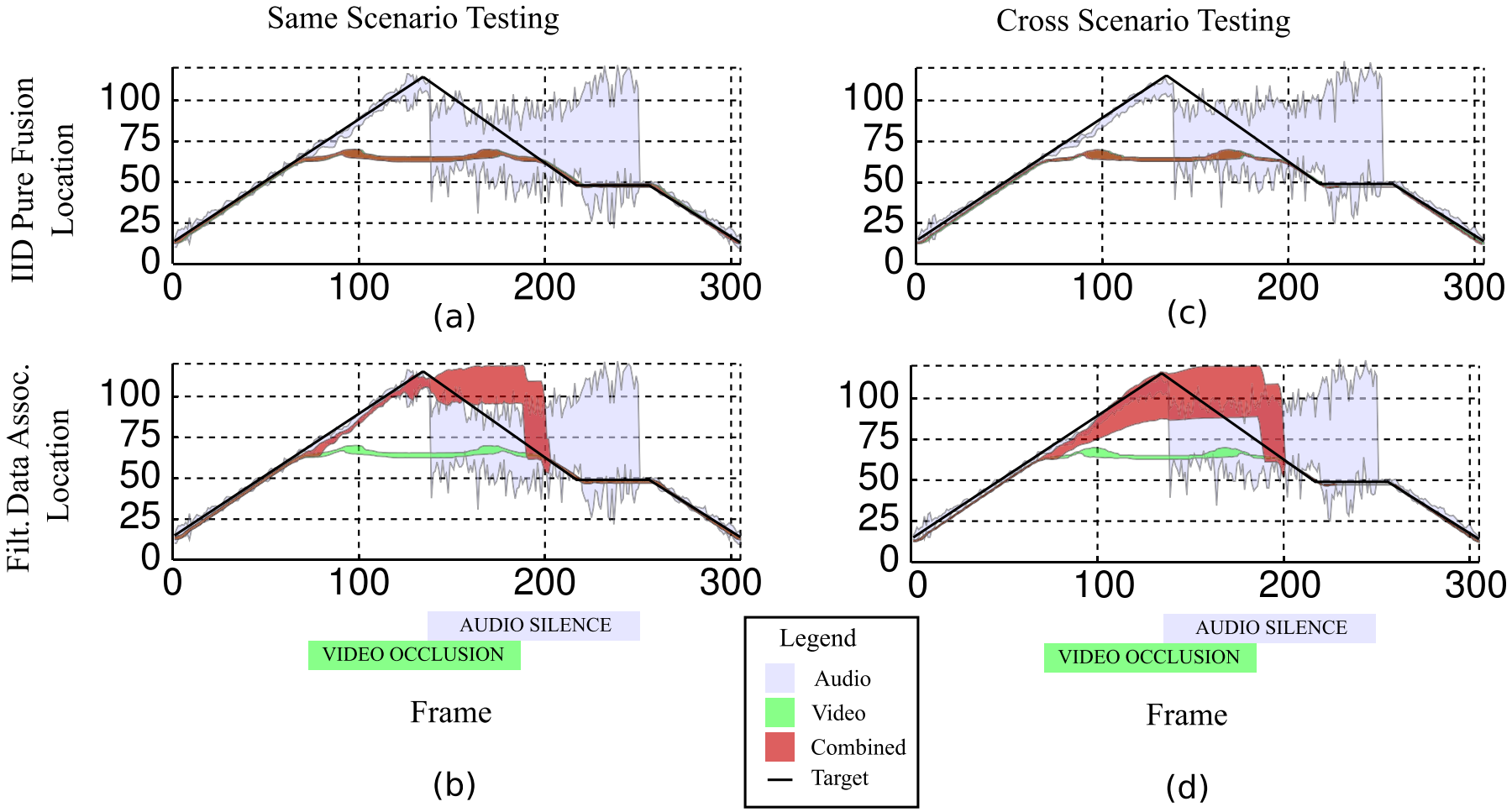


Figure 3.6: Performance of AV tracking of the mechanically controlled target. Input is 24 test sequences of different statistics (see text). Ground truth is indicated by the plain black line. The light/blue and medium/green shaded regions indicate the distribution of tracking estimates over all recorded sequences for audio and video only tracking respectively. The dark/red shaded region indicates distribution of of estimates for (a,c) IID pure fusion model and (b,d) filtered data association model. In (a,b) the models are tested on data of the same statistics for which they were trained. In (c,d) the base statistics are used to train the model which is then tested on data of all the other statistics. Shaded bars under the plots indicate the regions of video and occlusion and audio silence in the sequence.

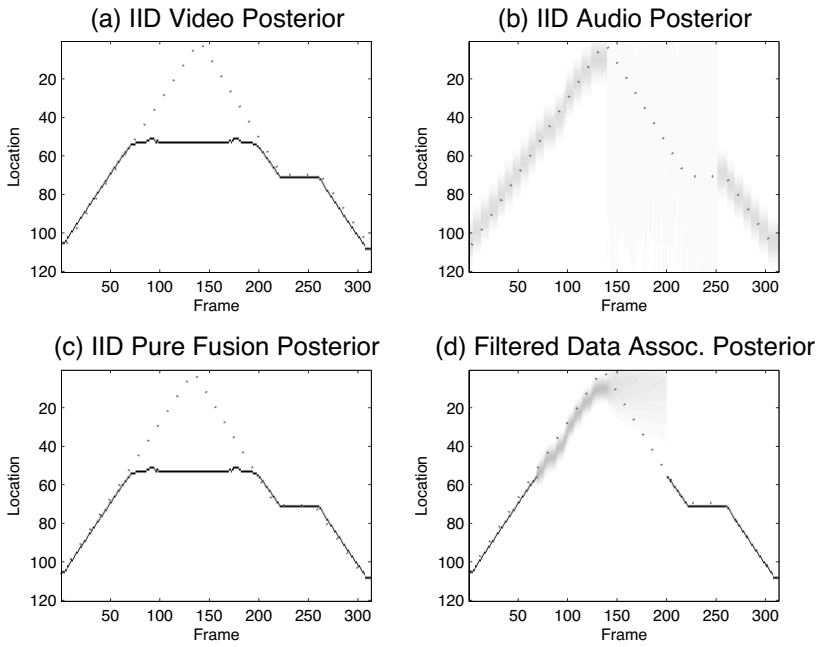


Figure 3.7: Posterior distribution over mechanical target location computed for a typical same condition trial. Distribution computed by (a) video input only, (b) audio input only, (c) IID pure fusion model, (d) filtered data association model. Ground truth is now indicated by the dashed line.

ant, we compute four performance measures:

1. Track percentage: the percentage of successfully tracked frames, defined as those for which the model output is within $\pm$10pixels of the true target location.

2. Accuracy: the average absolute error in pixels between the model's estimate and the true location for those frames in which the target was being tracked.

3. Audio detection rate (ADR): the percentage of frames for which the audio was correctly identified as being audible or not.

4. Video detection rate (VDR): the percentage of frames for which the video was correctly identified as being visible or not.

These are along the lines of standard evaluation measures for multimodal detection and tracking, e.g., as formalised by the CLEAR evaluation campaign ([cle, 2006], [Stiefelhagen and Garofolo, 2007]).

A key aim of multimodal perception is to improve performance over any individual modality, so ideally the combined models should perform better than the individual modalities. In this case, however, the pure fusion models [Beal et al., 2003] do not outperform the unimodal tracking under any measure because the fusion is dominated by the video, which can be unreliable during occlusion. In contrast, the models developed here – which try to infer the structure on the fly — generally succeed in doing so (Table 3.2, ADR, VDR columns), which allows them to fuse the modalities only when appropriate and track most reliably. In particular, the filtered data association model we develop here outperforms the IID pure fusion model developed in [Beal et al., 2003] by some margin (Table 3.2, bold). The performance reported in the left and right sections of Table 3.2 is for the same and cross condition testing respectively. As expected, the same condition performance is generally better than the cross condition performance for each measure. However, it is worth noting that perfect cross-condition detection performance is not unambiguously positive as eventually we will want to discriminate among different sources of different statistics during multi target tracking as we discuss in the next section.

3.3.2.3 Limitations of the Model

It is worth mentioning some limitations of the current model before continuing. Thus far we have discussed only one dimensional tracking in the horizontal plane. This is because most of the data we are interested in exhibits variability primarily in this plane and because the two element microphone array only provides information in this plane, rendering multimodal cue combination only interesting in this plane. Using the techniques in Section 3.2.4.1 ([Frey and Jojic, 2003]), it is simple and efficient to compute visual likelihoods in both axes. However, this would render the tracking Markov model as presented here unfeasibly slow requiring, for example, sparse matrix techniques such as [Jojic et al., 2000]. A stronger limitation is that the difficulty of representing visual rotation and scaling with TMG [Frey and Jojic, 2003] precludes tracking these variations efficiently in our parametric framework. However, to some extent the multimodal framework developed here can alleviate this problem as tracking (in the horizontal plane at least) can continue based on the audio modality even if vision fails due to excessive rotation or scaling.

Another class of potential problem relates to the unsupervised EM learning algorithm in the TMG framework [Jojic et al., 2000, Frey and Jojic, 2003] rather than the tracking procedure. In trying to find a single set of parameters θ that maximize the like-

	SAME				CROSS			
	Track %	Accuracy	ADR %	VDR %	Track %	Accuracy	ADR %	VDR %
Aud Only	72.3 ± 2.5	2.57 ± 0.08	-	-	71.8 ± 3.1	2.01 ± 0.08	-	-
Vid Only	65.6 ± 1.1	2.52 ± 0.01	-	-	65.5 ± 0.6	3.10 ± 0.01	-	-
PF IID	$\mathbf{65.6} \pm 1.1$	2.52 ± 0.01	-	-	$\mathbf{65.5} \pm 0.6$	3.10 ± 0.01	-	-
PF Filt	65.6 ± 1.1	2.52 ± 0.01	-	-	65.5 ± 0.6	3.10 ± 0.01	-	-
DA IID	81.5 ± 2.4	2.67 ± 0.01	96.7 ± 10.2	100 ± 0	77.7 ± 6.5	3.19 ± 0.05	91.4 ± 20	100 ± 0
DA Filt	$\mathbf{86.3} \pm 2.6$	2.70 ± 0.01	96.7 ± 10.2	100 ± 0	$\mathbf{83.2} \pm 6.6$	3.24 ± 0.06	91.4 ± 20	100 ± 0

Table 3.2: Quantitative evaluation of AV tracking results using mechanically controlled target. Results compare percentage of frames with (i) successful tracking, (ii) correct inference of audibility (ADR) and (iii) visibility (VDR) of target – only the last two methods computed ADR/VDR. For successfully tracked frames, accuracy of tracking in terms of pixel error is also shown. Table SAME indicates tests performed using input of the same statistics as the training data. Table CROSS indicates tests performed using one trained model and input of all the other different statistics. The model of [Beal et al., 2003] corresponds to the row PF IID. See text for detailed explanation of conditions.

lihood of the data $\theta = \text{argmax}_\theta p(\{\mathbf{x}_1^t, \mathbf{x}_2^t, \mathbf{y}^t\}_{t=1}^T | \theta)$, there may be many local maxima. For example, the foreground video model $(\mathrm{Z} = 1)$ can potentially learn a parameter μ to explain every video frame t as the stationary ($l^t = 0$) (true) room background, with the (true) foreground user being explained away by noise Ψ on every frame. This could be more likely than the intended maxima if:

1. The user's appearance area is very small compared to the background area.

2. The user is more frequently occluded than not in the training data.

3. The user is silent more frequently than not in the training data.

4. The actual background of the room is highly structured, making large translations difficult to explain by rotation as required in TMG [Jojic et al., 2000].

If many of these factors are true, inappropriate templates may be learnt and $\mathrm{Z}^t = 1$ may be inferred for all frames. Changing the parametric framework to one with a more explicit notion of layers [Williams and Titsias, 2004, Jojic and Frey, 2001] may be necessary to entirely avoid these problems.

3.3.3 Inference for Multiple Sources

We have seen the benefits of a principled probabilistic approach to data association for user detection, robust tracking through occlusion and multimodal user verification. However, the real value of explicit structural inference comes in multi-object scenarios where the question of single target user verification generalizes to the *who said what* problem (see Figure 3.8 for schematic). Exact inference unfortunately becomes exponentially more expensive in the maximum number of objects as the objects' states become conditionally dependent, given their shared observations. Moreover, it is difficult to represent this scenario properly within the TMG framework, as every observation element must be explained by the same source. Nevertheless, we shall see that with some small changes to the model as described in eq. (3.2) and Figure 3.2, we can efficiently approximate inference in the multi-target scenario and solve the who said what problem.

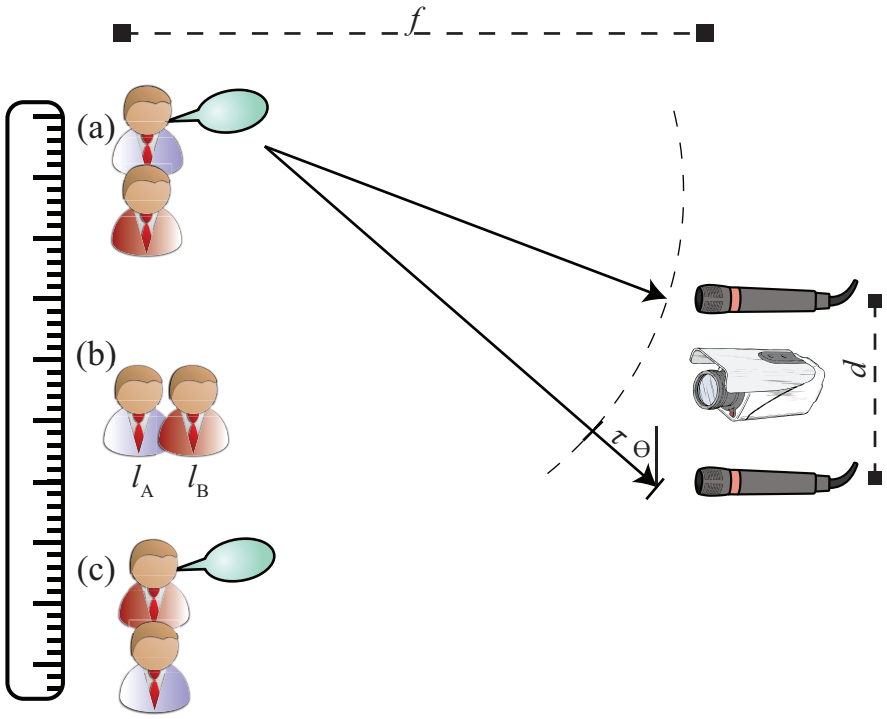


Figure 3.8: Schematic of scenario for multi-person audio-visual tracking and scene understanding. Participant locations l_A, l_B, visibility, and audibility are inferred on the basis of visual appearance and inter-microphone time delay τ.

3.3.3.1 Multi-target Tracking Framework

A first approximation to multi-target inference is simply to ignore the conditional dependency between the latent states of each object. Two separate instances of the model, (such as that of eq. (3.2) and Figure 3.2), can each can be initially trained with data containing a target of interest. In other words, data D_A containing samples of target A is used to train a Model A using EM with ML parameters $\theta_A = \text{argmax}_{\theta_A} p(D_A|\theta_A)$, and Model B learns the ML parameters θ_B from data D_B containing samples of target B, $\theta_B = \text{argmax}_{\theta_B} p(D_B|\theta_B)$. Once trained, these models can perform multi-target tracking and scene understanding by simultaneous but independent inference, such that Model A computes $p(l_A^t, \mathrm{W}_A^t, \mathrm{Z}_A^t | D^{1:t}, \theta_A)$ and Model B computes $p(l_B^t, \mathrm{W}_B^t, \mathrm{Z}_B^t | D^{1:t}, \theta_B)$. This is linear rather than exponential cost in the number of targets.

The suitability of this approach depends on the extent to which data from each target behaves like explainable noise from the perspective of the tracker concerned with the other target. This assumption does not quite hold given the model as introduced in Section 3.2 and trained as described in Section 3.3. The main reasoning

behind this is the fact that after learning, the parameters in θ describe two classes of audio data: "foreground" speech of large amplitude and associated source location and "background" office noise of smaller amplitude and uncorrelated source location. In the multi-target scenario, there are now three empirical classes of audio data: foreground associated speech (generated from the target of interest), foreground disassociated speech (generated from another target not of interest, and hence, with τ uncorrelated with l), and background office noise.

To decide if to associate a given frame of audio data $(\mathbf{x}_1, \mathbf{x}_2)^t$ with its target, Model A computes $p(\mathrm{W}_A^t | D^{1:t}, \theta_A) = \sum_{Z_A, l_A} p(l_A^t, \mathrm{W}_A^t, \mathrm{Z}_A^t | D^{1:t}, \theta_A)$. This depends on two important factors in the generative model: Firstly, the three way *match* between the peak of this likelihood as a function of l, the prior predicted location probability $p(l^t | D^{1:t-1})$ and the likelihood of the video observation $p(\mathbf{y}^t | l_A, \mathrm{Z}_A, \theta_A)$. (This is exactly the point that was introduced in Section 2.1, and illustrated in Figures 2.3 and 2.4.) Secondly, the association depends on the template match as specified by the the likelihood of the audio data under the background and foreground distributions unique to the audio-visual model:

$$\begin{aligned} p(\mathbf{x}_1^t, \mathbf{x}_2^t | l_A, \mathrm{W}_A, \theta_A) &= \int_{\mathbf{a}} \sum_{\tau} \mathcal{N}(\mathbf{x}_1 | \mathbf{a}, \upsilon_1)^w \mathcal{N}(\mathbf{x}_1 | \mathbf{0}, \sigma_1)^{\overline{w}} \\ &\quad \cdot \mathcal{N}(\mathbf{x}_2 | \mathbf{T}_\tau \mathbf{a}, \upsilon_2)^w \mathcal{N}(\mathbf{x}_2 | \mathbf{0}, \sigma_2)^{\overline{w}} \mathcal{N}(\tau | \alpha l + \beta, \omega). \quad (3.30) \end{aligned}$$

The new empirical class of data (disassociated speech) will be probable under the audio template likelihood model. At the same time, it will be unlikely in terms of the match between the *shape* of this likelihood, that of the video and predictive distribution from the Markov chain. In practice, this means disassociated speech would frequently be inappropriately classified as associated speech[5]. Therefore, we introduce a second background model to account properly for all three classes of audio data that are now present. Conveniently, the additional model only needs parameters already determined during learning. Let W now be three dimensional multinomial, defining the following three audio-modality likelihoods:

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2 | \mathbf{a}, \tau, \mathrm{W} = 1) &= \mathcal{N}(\mathbf{x}_1 | \mathbf{a}, \upsilon_1) \mathcal{N}(\mathbf{x}_2 | \mathbf{T}_\tau \mathbf{a}, \upsilon_2), \\ p(\tau | l, \mathrm{W} = 1) &= \mathcal{N}(\tau | \alpha l + \beta, \omega), \quad (3.31) \end{aligned}$$

[5]The observation likelihood under the background model $\overline{w}$ is very low as it is implausible that every component of the two 1000 dimensional background Gaussians $\mathcal{N}(\mathbf{x}_1 | \mathbf{0}, \sigma_1) \mathcal{N}(\mathbf{x}_2 | \mathbf{0}, \sigma_2)$ simultaneously become large. This is a much stronger effect than the mismatch in shape between the foreground likelihood and the video and predictive distributions, which occur only in one dimension.

$$
\begin{aligned}
p(\mathbf{x}_1, \mathbf{x}_2 | \mathbf{a}, \tau, \mathrm{W} = 2) &= \mathcal{N}(\mathbf{x}_1 | \mathbf{a}, \upsilon_1) \mathcal{N}(\mathbf{x}_2 | \mathbf{T}_\tau \mathbf{a}, \upsilon_2), \\
p(\tau | \mathrm{W} = 3) &= \mathcal{U}(\tau), \qquad (3.32) \\
p(\mathbf{x}_1, \mathbf{x}_2 | \mathbf{a}, \tau, \mathrm{W} = 3) &= \mathcal{N}(\mathbf{x}_1 | \mathbf{0}, \sigma_1) \mathcal{N}(\mathbf{x}_2 | \mathbf{0}, \sigma_2), \\
p(\tau | \mathrm{W} = 3) &= \mathcal{U}(\tau). \qquad (3.33)
\end{aligned}
$$

The foreground model $\mathrm{W} = 1$ is unchanged (eq. (3.31)), the first background model $\mathrm{W} = 2$ (eq. 3.32)) now accounts for signals with the statistics of speech but without any expected correlation with the predicted location or the likelihood of the video and the second background model $\mathrm{W} = 3$ (eq. (3.32)) is also unchanged from before, accounting for background office noise.

3.3.3.2 Multi-target Tracking: Detailed Example

The results for such a multi-target scenario (Schematic, Figure 3.8) are illustrated in Figure 3.9. In this scene, two users are having a discussion while moving around, occasionally passing in front of – and therefore occluding – each other (Figure 3.8(b)). The raw input waveform and video data are illustrated in Figure 3.9(a),(b). Model A and B have previously been trained independently on data (similar to that of Figure 3.3(a),(b)) containing their respective users and learnt – amongst other parameters $\theta_{A,B}$, the video templates $\mu_{A,B}$ shown in Figure 3.9(c),(d). The trained models each now report the posterior distribution over location and data association for their user u, $p(\mathrm{W}_u^t, \mathrm{Z}_u^t, l_u^t | D^{1:t}, \theta_u)$.

The smoothed posterior distribution over audio association $p(\mathrm{W}_u = 1 | D)$ is shown in Figure 3.9(g) with a light line for user A and a dark line for user B. The turn-taking behaviour in the conversation is clear with the alternating modes in the distribution for each. The posterior over video association $p(\mathrm{Z}_u | D)$ is shown in Figure 3.9(h). The initial presence of user A in the video is indicated by the initially high value for $p(\mathrm{Z}_A | D)$, and the subsequent entrance of user B is indicated by the rising initial value for $p(\mathrm{Z}_B | D)$. The fact that the subsequent occlusions as the users pass each other in the scene are correctly inferred is clear by the later dips in the line. Finally, the MAP location of each user is illustrated in Figure 3.9(i) along with the audio and video likelihood modes for each model. Similarly to the situation in Section 3.3.1, during visual occlusion, the video likelihood modes are quite spurious, but the detection and tracking functionality ensures the spurious modes are ignored until the user is visible again.

Model	Track %	Accuracy (Pixels)
AO	28.9 ± 7.0	4.33 ± 0.52
VO	86.3 ± 19.1	1.99 ± 1.06
PF IID	$\mathbf{86.3} \pm 19.1$	1.99 ± 1.06
PF Filt	86.4 ± 19.1	1.99 ± 1.06
DA IID	86.2 ± 15.2	2.01 ± 1.07
DA Filt	$\mathbf{88.7} \pm 12.6$	2.04 ± 1.14

Table 3.3: Summary of multi-user tracking performance. Track % indicates percentage of time the tracker's output was on target — within ± 10 pixels of the true target location. Accuracy indicates the absolute error in pixels of the tracker for the correctly tracked frames.

An important and novel feature of this framework is that segmentation of the original raw speech data $\mathbf{x}_1$,$\mathbf{x}_2$ is now provided – as a byproduct of inference, by the posterior probability of audio association $p(\mathrm{W}_u|D)$. That is, $p(\mathrm{W}_u^t = 1|D^{1:t})$ defines the posterior probability that the speech at time t $(\mathbf{x}_1^t, \mathbf{x}_2^t)$ originated from user u. It is, therefore, the probabilistic answer to the question of who uttered the current frame of speech. The speech segments uttered by each user are extracted from the raw data using $p(\mathrm{W}_u^t = 1|D^{1:t})$, and illustrated in Figure 3.9(e),(f). This is the solution to the *who said what* problem. In contexts such as conversation understanding, transcription and summarization [Hain et al., 2005], the segmented speech signals could then be passed on to a speech processing system to produce a speaker labelled transcription.

3.3.3.3 Multi-target Tracking: Quantitative Evaluation

In this section, we summarize the quantitative performance of the models in a multi-target tracking context. We recorded five multi-party conversation video sequences of approximately one minute each along the lines of the one examined in detail in Section 3.3.3.2. The sequences included some different room configurations and users – this necessitated the learning of the different audio-visual appearances. To create ground truth, we manually labelled the location, visibility and speaking status of the users in each frame. Given the ground truth data, we were able to quantify performance using a similar procedure to that described in Section 3.3.2.2.

Table 3.3 details the tracking performance of the models in this multi-target scenario averaged over all the recorded sequences. Based on the key measure of percent-

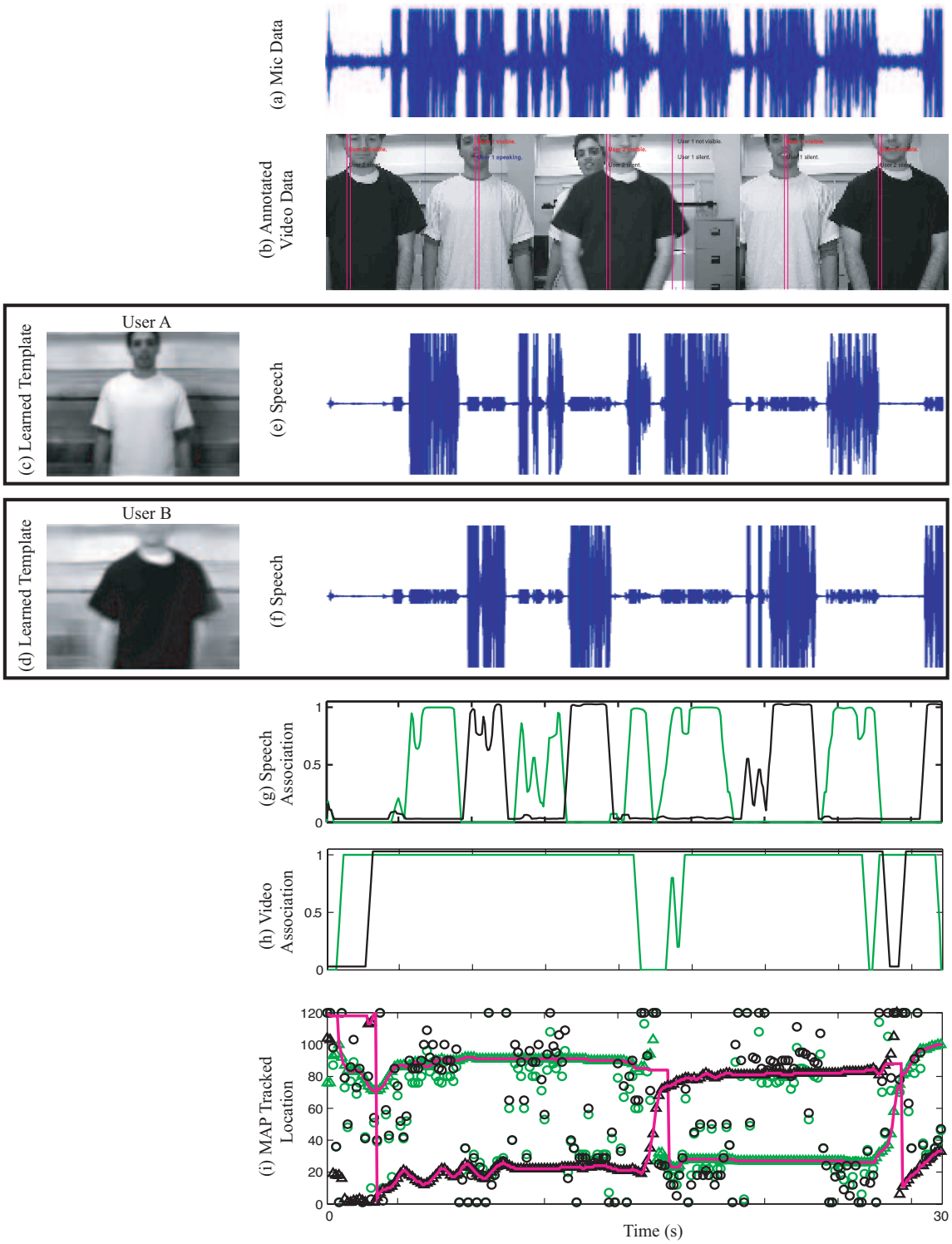


Figure 3.9: AV multi-object tracking and scene understanding results. (a) Raw audio data and (b) sample video frames from a sequence where two users are conversing and moving around, occasionally occluding each other. (c,d) Learnt templates for the two users. (e,f) Speech segments inferred to belong to each user. Posterior probability of audibility (g) and visibility (h) for user A (light/green) and B (black). (i) Multi user tracking. Audio likelihood peaks are shown as circles and video likelihood peaks as triangles. MAP locations are shown by the two dark lines.

	ADR	VDR
Total Error	$19.0 \pm 8.5\%$	-
Effective Error	$12.9 \pm 7.5\%$	3.0 ± 4.5 %
False Pos	$4.7 \pm 6.2\%$	$1.3 \pm 4.0\%$
False Neg	$8.1 \pm 2.7\%$	$1.7 \pm 2.9\%$

Table 3.4: Individual user detection performance in a multi-party scenario.

		Actual		
		User A	User B	None
Reported	User A	**72.6**$\pm$16.0%	$9.5 \pm 15.0\%$	5.5$\pm$5.3%
	User B	$11.0 \pm 13.4\%$	**74.6**$\pm$21.7%	3.3$\pm$3.3%
	None	$16.4 \pm 8.6\%$	$15.9 \pm 9.3\%$	**91.2**$\pm$7.6%

Table 3.5: Confusion matrix for multi-user speech segmentation.

age of successfully tracked frames (Track %), the audio-only tracking performance is much lower than that of Table 3.2. This is because the speech signal is less precisely localizable and more intermittent than the noise signal used in Section 3.3.2.2. Nevertheless, combining the audio and video modalities with structure inference allows the filtered data association model to perform better than either modality alone, as well as better than the pure fusion model [Beal et al., 2003].

Next we evaluate the audio-visual association performance. The earlier model variants do not compute this, so we focus on the performance of the final filtered data association model. The audio model now has three possible structures W. The total error can first be computed as the percentage of frames for which the ground truth W^{gt} and the model's MAP estimate W^{est} do not match, $\mathrm{W}_u^{gt} \neq \mathrm{W}_u^{est}$. Since we are mostly interested in detecting the correct speaker $\mathrm{W}_u = 1$ and not the nature of the negatives ($\mathrm{W}_u = 2$ vs $\mathrm{W}_u = 3$), we combine the two negative categories when computing the effective error rate. The effective error rate can then be further broken down into false positives, i.e. reporting user u is speaking when actually he is silent ($\mathrm{W}_u = 1$ but $\mathrm{W}_u^{gt} = 2, 3$) and false negatives, i.e. reporting user u is silent when actually he is speaking ($\mathrm{W}_u = 2, 3$ but $\mathrm{W}_u^{gt} = 1$). The detection rates are computed similarly for the video modality. The results are reported in Table 3.4.

In this multi-party conversation context, an interesting quantity is the accuracy with which the model can assign speech segments to the users. Therefore, in Table 3.5, we

	Hospedales	Gatica-Perez	Checka
Track %	89%	99%	Not Quantified
Speaker Detection %	80%	87%	80%

Table 3.6: Tracking and association performance: comparison to other results in the literature.

also report the average confusion matrix between the actual and reported speaker of each segment in terms of containing speech from user A, user B or neither. The model performs well, correctly assigning at least 72% of the speech segments to the user uttering them.

3.3.3.4 Multi-target Tracking and association: Comparison to Related Research

While there are have been numerous studies on AV sensor fusion for tracking and for speaker association, there have been relatively few addressing the simultaneous solution of these problems, as we have investigated in this chapter ([Hospedales and Vijayakumar, 2008]). We are aware of two studies which tried to solve similar tasks to ours - those of Gatica-Perez et al. [Gatica-Perez et al., 2007] and Checka et al. [Checka et al., 2004]. Table 3.6 reports the overall average percentage of correctly tracked frames and correctly labelled speaker assignments in each of these studies. Note, however, that comparison of the quantitative results is not possible, as the specific data-sets, problem constraints, and evaluation criteria (e.g., definition of a tracking "hit") vary.

We now contrast the problem scenarios and models. Both these studies used particle filtering for inference, which permitted quite flexible modelling of the task (e.g., AV mapping). In contrast, our approach was to some extent constrained by the need to construct a model with analytically tractable inference. This required the use of various additional approximations, such as linearity in the time-delay to location AV mapping (eq. (3.1)). In principle, learning the AV mapping parameters from extensive data sets should allow better performance (as well as more convenience) than limited manual calibration; however, in practice, the poverty of the linear mapping limited the available performance gains here.

The study of Checka et al. ([Checka et al., 2004]) performed tracking in two dimensions plus scale, as opposed to our azimuthal, fixed scale tracking. They applied much more powerful sensor capabilities than ours: two cameras and a 16 element

microphone array. To simplify their model, they only inferred speaking status (i.e., our audio association, W) without also computing visibility. This precluded tracking through visual occlusion, which we achieved and exploited. Their particle filter model required extensive pre-calibration to specify the AV mapping and the image for background subtraction etc. Importantly, they constrain their problem to tracking human figures (modelled as cylinders) and exploit this knowledge in their visual appearance model to improve tracking accuracy. In contrast, our appearance model is a generic learned template, which makes it more broadly applicable (e.g., for tracking vehicles in surveillance [Jojic et al., 2000]).

The study of Gatica-Perez et al. ([Gatica-Perez et al., 2007]) performed tracking and association for up to three users, in two dimensions plus scale. They also employed a more powerful (8 element) microphone array and extensive AV pre-calibration. In this case, they applied strong domain specific knowledge: constraining their model to specifically track human faces. This allowed them to exploit skin color and facial feature/contour detection in tracking. They also only infer speaking status (and not visibility), again precluding tracking through video occlusion.

3.3.4 Summary

In this section, we have illustrated the application of the ideas introduced in Sections 2.1 and 3.2 to a real audio-visual scene understanding problem. Multisensory detection, verification and robust tracking through occlusion of either or both modalities is achieved through inference of latent state and structure. The data association inference turns out to depend on a combination of three effects: correlation between the shape of the observation likelihoods in each modality; correlation between the shape of the observation likelihoods and the predictive distribution; and the goodness of the template match in each modality.

The multi-target data association problem is more interesting as the solution to it represents explicit relational knowledge of who was present (visible) when and who said what when. While expensive to compute exactly, in this application, an independence approximation in which the background models for each user explain data generated by the other user turns out to be sufficient for robust multi-target tracking and data association. A probabilistic segmentation of the speech is achieved as a byproduct of the explicit computation of data association.

3.4 Discussion

In this chapter, we introduced a principled formulation of multisensor perception and tracking in the framework of Bayesian inference and model selection in probabilistic graphical models. Pure fusion multisensor models have previously been applied in machine perception applications and in understanding human perception. However, for sensor combination with real world data, extra inference in the form of data association is necessary, as most pairs of signals should not actually be fused. Moreover, in many cases, inferring data association is in itself an important goal for understanding structure in the data. For example, a speech transcription model should not associate nearby background speech of poorly matching template and uncorrelated spatial location with the visible user when he is silent. More significantly, to understand a multi-party conversation, the speech segments need to be correctly associated with person identity. In our application, the model computes which observations arise from which sources by explicitly inferring association, so it could for example, start a recording when the user enters the scene or begins speaking and segment the speech in a multi-party conversation.

3.4.0.1 Related Research

While we have discussed relevant previous research in Section 1.1 and Section 3.3.3.4, it is worth contrasting our study against some other broadly related pieces of recent and ongoing work. In radar tracking and association, some work [Stone et al., 1999] uses similar techniques to ours; however, popular methods [Bar-Shalom et al., 2005] tend to be more heuristic, necessarily use stronger assumptions and approximations (e.g., Gaussian posteriors) and use highly pre-processed point-input data. One interesting contrast between these candidate detection based approaches and our generative model approach is that we avoid the expensive within-modality data association problem typical of radar. This also enables use of signature or template information in a unified way along with cross-modality correlation during inference, which is exploited to good effect in our AV application.

Structure inference issues also arise in some other very different fields, such as scientific citation indexing [Pasula et al., 2003]. Here, sources (scientific papers) induce observations (particular citations). Because the particular string used for the citation of a given paper varies from instance to instance, this observation process is noisy. Then asking, for example, whether two given citations refer to the same specific paper also

poses a structure inference problem, and requires related techniques to solve.

In audio-visual processing, [Siracusa and Fisher, 2007] independently proposes a model which computes association between two speakers and their speech segments by inferring the presence or absence of conditional dependencies. However, this model it specific to the association task and does not handle the full tracking and AV template learning problem which we address simultaneously here.

In computer vision, [Williams and Titsias, 2004] and [Jojic and Frey, 2001] describe techniques related to ours for unsupervised learning and tracking of multiple objects in video using greedy and variational inference approximations, respectively. These do not require the independent learning for each target used in our framework. However, in using only one modality, [Williams and Titsias, 2004, Jojic and Frey, 2001] avoid the multimodal data association problem which we address here.

3.4.0.2 Future work

Our work as described here generalizes existing pure fusion models and, using a single probabilistic framework, provides a principled solution to questions of sensor combination including signature, fusion, fission and association. As our AV application illustrates, computing the exact posterior over source state and multi-target data association for real problems is potentially even real-time.

Performance could potentially be improved by adding support for color video to the model. This would not add significant computational overhead, but would make it easier to disambiguate multiple users. As mentioned in Section 3.3.3.4, it could also be beneficial to compute and use more pre-processed audio and video features as input. This could be in addition to, or instead of modelling every pixel and audio-sample directly as we do now. Our more general theme for future research is to close the sensorimotor loop by integrating our existing work on sensorimotor control [Vijayakumar et al., 2001] with these probabilistic perceptual models, to extend them into the domain of active perception.

Chapter 4

Bayesian Structure Inference in Human Perception

In this thesis, we have argued for a structure inference interpretation of multisensory perception in the presence of ambiguous data association. This was on the grounds of both robustness in state inference and the value of explicit knowledge of data association. In the previous chapter, we applied our structure inference approach to develop a machine perception system for learning an audio-visual tracking and data association task, where the structure inference allowed tracking through occlusion and conversation segmentation.

What of human perception? The human perceptual system potentially faces a variety of similar problems (e.g., intermittent observations and the need to know who said what in multi-party conversations) to those solved by our machine learning system in Chapter 3. An interesting question then, is whether the human perceptual system makes use of any similar computations? We explore this question in this chapter.

As we will see, many recent experiments in human psychophysics have – intentionally or inadvertently – presented stimuli of uncertain data association, and found that classical sensor fusion theory fails to explain the results. We will attempt to address these explanatory shortcomings with our structure inference approach to multisensory perception. We introduce probabilistic modelling for psychophysics in Section 4.1. We then apply our techniques to understanding experiments (similar to those of Chapter 3) in audio-visual localization and association in Section 4.2. In Section 4.3, we apply our techniques to understanding another set of experiments in the completely different paradigm of oddity detection, and with completely different modalities: visual-haptic size cues and stereo-texture slant cues. Finally, we summarize the commonality of our

contributions in these domains, and their relation to other research in Section 4.3.4.

4.1 Modelling Human Perception

Bayesian ideal observer modelling is an elegant and successful approach to understanding human perception [Kersten et al., 2004]. Many recent studies have applied it to understand multisensory fusion in human perception across a variety of combinations of modalities [Alais and Burr, 2004, Ernst and Banks, 2002, Jacobs, 1999, van Beers et al., 2002, Ernst and Bulthoff, 2004, Gepshtein and Banks, 2003, Landy and Kojima, 2001, Battaglia et al., 2003]. In this approach, a generative probabilistic model for the perceptual process is defined. This describes the way in which signals are generated by a source, and how they are then observed — including any distorting noise processes. This is analogous to the generative modelling approach commonly taken in machine learning (e.g., Section 3.2.1), although simpler parametric forms tend to be used. Predictions made by the results of inference in this model can then be compared to experimental results.

In the next two sections, we review standard ideal observer models for sensor fusion in psychophysics and the experimental designs used to test them. This will provide the context for the subsequent discussion of structure uncertainty in psychophysics.

4.1.1 Ideal Observer Modelling for Sensor Fusion

As we saw in Section 1.1, classical sensor fusion theory assumes that multisensory observations x_i in modalities i are generated from some source y in the world, subject to independent noise in the environment and physical sensor apparatus, e.g., $x_i \sim \mathcal{N}(y, \sigma_i^2)$. The sensors may have different variances σ_i^2.

For example, in [Ernst and Banks, 2002], subjects make haptic x_h and visual x_v observations of a bar's height y, and must report their combined estimate $(\hat{y}_{h,v})$ of the true height. This is an inference problem which can be represented by a generative graphical model shown in Figure 4.1. Under this particular noise model, the posterior distribution of the height estimate is a Gaussian: $p(y|x_h, x_v; \sigma_h^2, \sigma_v^2) = \mathcal{N}(y; \mu_{y|h,v}, \sigma_{y|h,v}^2)$, with mean and variance given by eqs. (4.1) and (4.2)[1]:

[1] See Appendix A.1.2.3 for details and derivation.

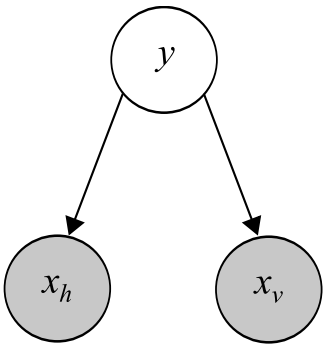


Figure 4.1: Classical sensor fusion model. Bar size y is inferred on the basis of visual and haptic observations x_v and x_h [Ernst and Banks, 2002].

$$\hat{y}_{h,v} = \mu_{y|h,v} = \frac{\sigma_h^{-2}}{\sigma_h^{-2}+\sigma_v^{-2}} x_h + \frac{\sigma_v^{-2}}{\sigma_h^{-2}+\sigma_v^{-2}} x_v, \tag{4.1}$$

$$\sigma^2_{y|h,v} = \frac{\sigma_v^2 \sigma_h^2}{\sigma_v^2+\sigma_h^2}. \tag{4.2}$$

For this Gaussian posterior $p(y|x_h,x_v)$, the optimal estimate to make ($\hat{y}_{h,v}$) under the standard mean square cost function [Kording and Wolpert, 2004b] is actually the mean of the posterior eq. (4.1), which turns out to be the precision weighted mean of the individual observations .

Psychophysics experiments such as [Ernst and Banks, 2002, Alais and Burr, 2004] typically test human multisensory perception for optimality and the match to the ideal observer behaviour in two ways. Firstly, the variance of the optimal response $\sigma^2_{y|h,v}$ is less than the variance of the individual observations σ_v^2 and σ_h^2 eq. (4.2). Therefore,

- the distribution of a human's responses $\hat{y}_{h,v}$ to a multisensory stimulus should have a lower variance than their responses $\hat{y}_h, \hat{y}_v$ to the unimodal stimuli.

Secondly, the multisensory response of the ideal observer is the precision weighted mean of the unimodal observations eq. (4.1). Therefore,

- experimentally manipulating the variances σ_h^2, σ_v^2 of the individual modalities should produce the appropriate changes in the human perceptual response $\hat{y}_{h,v}$.

(The larger the ratio σ_h^2/σ_v^2 , the closer $\hat{y}_{h,v}$ will be to x_v and vice-versa.)

4.1.2 Ideal Observer Model Parametrisation

Before ideal observer models can be tested as described above, the model parameters must be estimated. In Chapter 3, our audio-visual model was able to learn these parameters directly from the data using EM. However, for the purpose of testing models of how humans combine multisensory information, we need to know the parameters representing the characteristics of the perceptual system.

Typically, maximum likelihood estimates of the variances σ_i^2 of individual modalities i are made separately using unimodal stimuli. In some designs, for repeated presentations of the unimodal stimulus, e.g., y_v, subjects report their specific estimate $\hat{y}_v$ of the stimulus, and the distribution of $\hat{y}_v$s over trials can be used to estimate σ_v^2 by maximum likelihood. For example, in recent audio-visual spatial localization experiments [Hairston et al., 2003, Wallace et al., 2004], σ_a^2 and σ_v^2 are determined as subjects point to the perceived locations of the unimodally presented audio and visual stimuli respectively.

In forced choice experimental designs, the procedure described above, wherein subjects directly report their best estimate of the observed signal is not suitable. For example, [Ernst and Banks, 2002] investigates the fusion of visual and haptic modalities in estimating the height of a bar. Rather than asking subjects to report the perceived height directly, subjects are asked to compare the heights of two bars and report the larger bar. In this case, if the bars are observed to have haptic heights $x_{h,1} \sim \mathcal{N}(y_{h,1}, \sigma_h^2)$ and $x_{h,2} \sim \mathcal{N}(y_{h,2}, \sigma_h^2)$, then they will sometimes correctly report the larger bar and sometimes not depending on the actual difference in heights and the noise on $x_{h,i}$ in a given trial. The distribution of responses can then be modelled by a cumulative Gaussian distribution, $p(x_{h,1} > x_{h,2} | y_{h,1}, y_{h,2}; \sigma_h^2) = p(x_{h,1} - x_{h,2} > 0 | y_{h,1}, y_{h,2}; \sigma_h^2)$. By fitting the observed distribution of responses to this model distribution, the observation noise parameter σ_h^2 is determined indirectly in contrast to the direct estimation method discussed above. Once the unimodal parameters are estimated, predictions for multisensory observations can be made (as discussed in Section 4.1.1) and tested.

4.2 Audio-Visual Localization

In this section, we discuss psychophysical models of audio-visual localization. This is the same problem that was solved by our machine learning system in Chapter 3, so we expect that the same theoretical framework should provide a good model of the

experiment.

A key phenomena in human audio-visual localization is the *ventriloquist effect*, wherein the apparent source of speaker's voice is largely determined by vision if available [Witten and Knudsen, 2005]. We are familiar with this in the every-day activity of television watching, where we perceive voices as coming from the on-screen characters rather than the speakers. Historically, this type of effect was explained by theories of visual capture, which suggested that for multisensory spatial localization tasks, vision dominated human perception [Witten and Knudsen, 2005].

More recently, this has been understood as a specific case of optimal sensor fusion [Alais and Burr, 2004, Witten and Knudsen, 2005]. Since the visual system enjoys very high spatial localization acuity compared to the auditory system, it should be expected to dominate during multisensory localization as we saw by the weighted mean in eq. (4.1). [Alais and Burr, 2004] reported the results of a multisensory localization task under experimental intervention which had the effect of reducing the spatial acuity of the subjects' visual system. There was to a gradual change in perceived location from the true visual location to the true auditory location as a function of the relative precision of the auditory and visual modalities. This led to the other extreme where, when visual precision was degraded strongly compared to auditory precision, an "inverse" ventriloquist effect was observed, in which the percept was dominated by audition. Hence we can understand the ventriloquist effect not as the result of visual capture specifically, but as a special case of optimal sensor fusion in which the highly unequal spatial localization precision of unmodified vision and audition result in visual dominance.

Interestingly, we saw a similar effect with the simplest audio-visual localisation model in Chapter 3 (see Section 3.3.1.1). When there was discrepancy between the visual and auditory modalities, the model's final fused percept tended toward the visual, rather than auditory likelihood peak. This was because, similarly to the case of human perception, the spatial precision of the model's visual observation was much higher than that of the auditory observation after EM learning. This was so despite the model having a completely different parametric form (discrete histogram) to that of the standard models of human perception (Gaussian).

Once updated to take structure uncertainty into account, our machine learning model for audio-visual localization managed to "see through" the ventriloquist effect illusion beyond a certain amount of discrepancy (Section. 3.3). Is the same thing possible for humans? A recent series of experiments [Hairston et al., 2003,

Wallace et al., 2004] investigated just this.

4.2.1 Experimental Background

Wallace et al. report the results of a series of audio-visual localisation and unity perception tasks [Hairston et al., 2003, Wallace et al., 2004]. In these experiments, auditory and visual stimuli were presented at a variety of locations (y_a and y_v, respectively) around the subject's periphery (see Figure 4.2 for schematic). These locations were sometimes discrepant (Figure 4.2(a) and sometimes coincident (Figure 4.2(b)). Subjects were required to report the specific perceived location of the *auditory* stimulus, and whether on not they perceived the auditory stimulus as being unified with the visual stimulus[2]. These studies quantified how the presence and location of the visual stimulus relative to the auditory stimulus affected the perception of unity and the perceived auditory location.

The experiments produced various striking results (see Section 4.2.3 and Figures 4.4 and 4.6 for details), notably:

1. The strength of the effect of the visual stimulus on the ultimate auditory location estimate was strongly correlated with the percept of unity or not. Moreover, in the event that non-unity was perceived, the final percept was repelled *away* from the visual stimulus.

2. The standard deviation of the estimate was strongly dependent on whether unity was perceived or not.

3. The localization error was strongly dependent on the report of unity or not. The unified trials had tightly and unimodally distributed error, and the non-unified trials had more widely distributed error.

None of these results can be explained within the framework of classical sensor fusion theory, (in any case, it has no notion of unity of percept or not). However, if the brain is indeed performing Bayesian inference to solve this problem — as assumed by the ideal observer theory — then, in addition to performing inference to localize the stimuli source, it may also be performing – or approximating – model selection:

[2]The discrepancies presented here extended to those large enough to allow the possibility of non-unified perception. This is in contrast to typical sensory integration experiments, e.g., [Alais and Burr, 2004], where the discrepancies were smaller and participants were explicitly instructed that the multisensory stimuli constituted unified events.

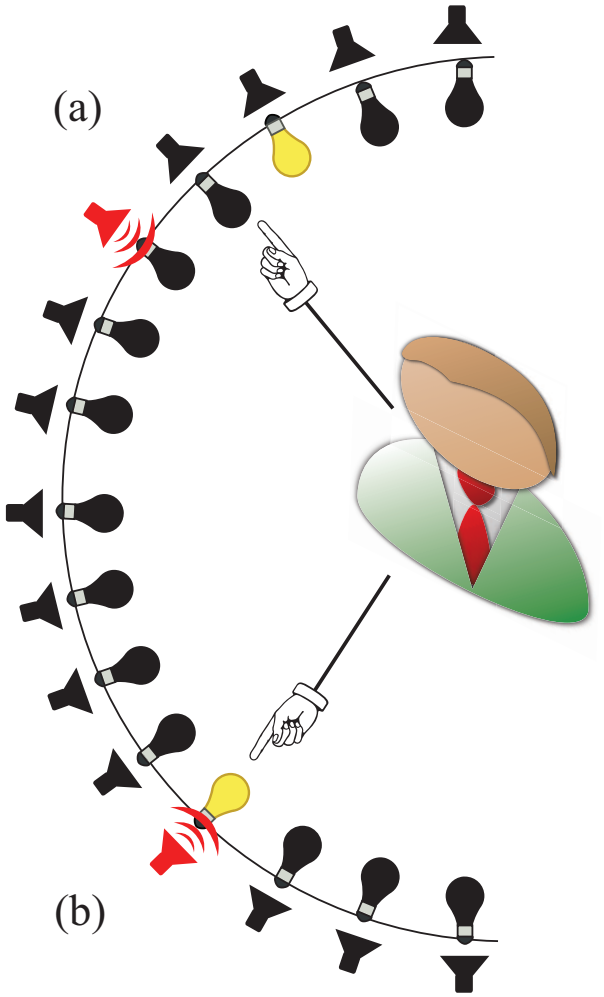


Figure 4.2: Schematic of the audio-visual perception task from [Wallace et al., 2004]. Participants observed audio-visual stimuli at a variety of (a) discrepant and (b) coincident spatial locations. They then reported whether they perceived the stimuli as unified or non-unified as well as the perceived location of the auditory stimulus.

to work out the relative likelihoods of a unified or non-unified explanation of the observations (e.g., as described in Section 2.2). We can test this more general theory of perception by comparing the model's predictions to the experimental results reported in [Wallace et al., 2004].

4.2.2 Modelling Audio-Visual Localization and Unity

4.2.2.1 Model

To model this experiment probabilistically — including the structure uncertainty — we use the approach introduced via the toy model in Section. 2.2. The graphical model in Figure 4.3 represents the full generative model for this experiment. Here the classical sensor fusion model (Figure 4.3, left) is included as a special case when the observations x_a and x_v are known to be unified (U = 1): both caused by the same stimulus y_a. Alternately, the observations may be uncorrelated (Figure 4.3, right, U = 0): x_a and x_v related to separate stimuli y_a and y_v. Subjects directly report both their best estimates of audio location $\hat{y}_a$ and unity $\hat{\mathrm{U}}$. To parametrize the model tractably, we assume that U is Bernoulli: $p(\mathrm{U}) = p_u^{\mathrm{U}}(1-p_u)^{(1-\mathrm{U})}$, and that all the other variables are Gaussian:

$$\begin{aligned} p(x_a|y_a,u) &= \mathcal{N}(y_a,p_a), \\ p(x_v|y_a,u) &= \mathcal{N}(y_a,p_v), \\ p(x_a|y_a,\overline{u}) &= \mathcal{N}(y_a,p_a), \\ p(x_v|y_v,\overline{u}) &= \mathcal{N}(y_v,p_v), \\ p(y_a) &= \mathcal{N}(\mu_y,p_y), \\ p(y_v) &= \mathcal{N}(\mu_y,p_y). \end{aligned}$$

We can estimate the parameters of this model by setting the audio and visual observation precisions (p_a and p_v) to their maximum likelihood values computed from the unimodal experiments reported in [Hairston et al., 2003][3]. The prior probability of unity p_u and parameter p_y representing the subject's prior belief about the stimulus locations cannot similarly be directly determined by the experimenter or modeller. These free parameters can be fixed heuristically (e.g., as uninformative, as we do here), or fit to the data (as in [Kording et al., 2007]).

[3]However, we follow [Kording et al., 2007], and assume that the visual unimodal responses were dominated by motor noise of $\sigma_m = 2.5$deg. So $\sigma_v = 0.01$deg and $\sigma_a = 7.6$deg.

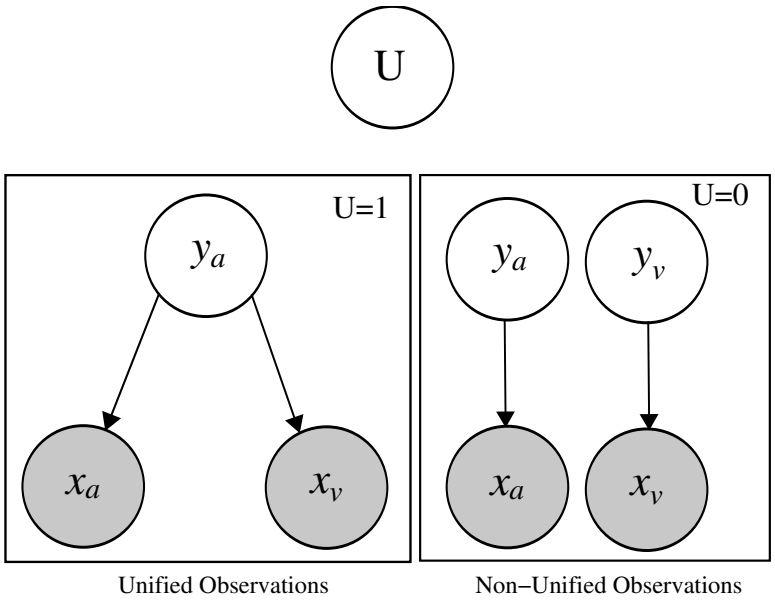


Figure 4.3: Graphical models to represent the audio-visual perception experiments of [Wallace et al., 2004]. A unified ($U = 1$) stimulus means that one latent source produces both observations. A non-unified ($U = 0$) stimulus means that the visual observation is produced independently of the auditory stimulus. Subjects are asked to report their percept of audio location y_a and stimulus unity U given audio and visual stimuli, x_a and x_v.

Note that it is not important whether y_v is explicitly represented as a latent variable or not, since it is never requested of the subjects. For the purposes of the model data likelihood $p(x_a, x_v|\mathrm{U})$, we could just as easily directly use a simple "background" distribution over x_v in the disassociated case, $p(x_v|\overline{u}) = \int p(x_v|y_v,\overline{u})p(y_v)dy_v$ as we did in the audio-visual application (Section 3.2.1, Figure 3.2).

4.2.2.2 Inference

To perform structure inference for this problem, the ideal observer computes the model posterior and reports the best estimate model $\hat{u} = \mathrm{argmax}_u p(u|x_a, x_v)$. By integrating the latent state, and assuming a uniform prior over models $p_u = 0.5$, we can compute the model posterior as follows[4]:

$$p(x_v, x_a|u) = \int p(x_a|y_a, u)p(x_v|y_a, u)p(y_a)dy_a, \tag{4.3}$$

[4]See Appendix A.1.2 for derivation and details

$$p(u|x_v,x_a) \propto \sqrt{\frac{p_a p_v p_y}{(2\pi)^2(p_a+p_v+p_y)}} \cdot \exp\left(-\frac{1}{2}\frac{(x_v-x_a)^2 p_a p_v + (x_v-\mu_y)^2 p_v p_y + (x_a-\mu_y)^2 p_a p_y}{p_a+p_v+p_y}\right), \quad (4.4)$$

$$p(x_v,x_a|\overline{u}) = \int\int p(x_a|y_a,\overline{u})p(x_v|y_v,\overline{u})p(y_a)p(y_v)dy_a dy_v, \quad (4.5)$$

$$p(\overline{u}|x_v,x_a) \propto \sqrt{\frac{p_v p_y}{2\pi(p_v+p_y)}}\exp\left(-\frac{1}{2}(x_v-\mu_y)^2(p_v^{-1}+p_y^{-1})^{-1}\right) \cdot\sqrt{\frac{p_a p_y}{2\pi(p_a+p_y)}}\exp\left(-\frac{1}{2}(x_a-\mu_y)^2(p_a^{-1}+p_y^{-1})^{-1}\right). \quad (4.6)$$

Given the model posteriors in eqs. (4.4) and (4.6), we can easily compute the location posterior posterior $p(y_a|x_a,x_v)$ as follows:

$$p(y_a|x_a,x_v) = \sum_{\mathrm{U}} p(y_a|\mathrm{U},x_a,x_v)p(\mathrm{U}|x_a,x_v). \quad (4.7)$$

This is in general a mixture of unity-conditional Gaussian posteriors $p(y_a|U,x_a,x_v)$, with one Gaussian for each of the hypotheses U about unity, as we saw in Chapter 2. The best estimate to report $\hat{y}_a$ depends on the loss function used, but we can assume the mean is reported under the typical mean squared error loss function [Kording and Wolpert, 2004b]. In this case the report is given specifically by the mean of eq. (4.7), which is:

$$\hat{y}_a = \frac{p_a x_a + p_v x_v + p_y \mu_y}{p_a+p_v+p_y}p(u|x_a,x_v) + \frac{p_a x_a + p_y \mu_y}{p_a+p_y}p(\overline{u}|x_a,x_v). \quad (4.8)$$

Given the derived inference equations (eqs. (4.4), (4.6) and (4.8)), we can compare the predictions of the model with the experimental results, as is discussed in the next section.

4.2.3 Results

4.2.3.1 Audio-Visual Bias

A key quantitative evaluation measure reported by [Wallace et al., 2004] was the audio-visual *bias*. This was defined as the amount that the estimated auditory location $\hat{y}_a$ deviated from the true auditory location y_a due to the presence of the visual stimulus y_v. Specifically, the bias B was defined as:

$$B = \frac{\hat{y}_a - y_a}{y_v - y_a}. \quad (4.9)$$

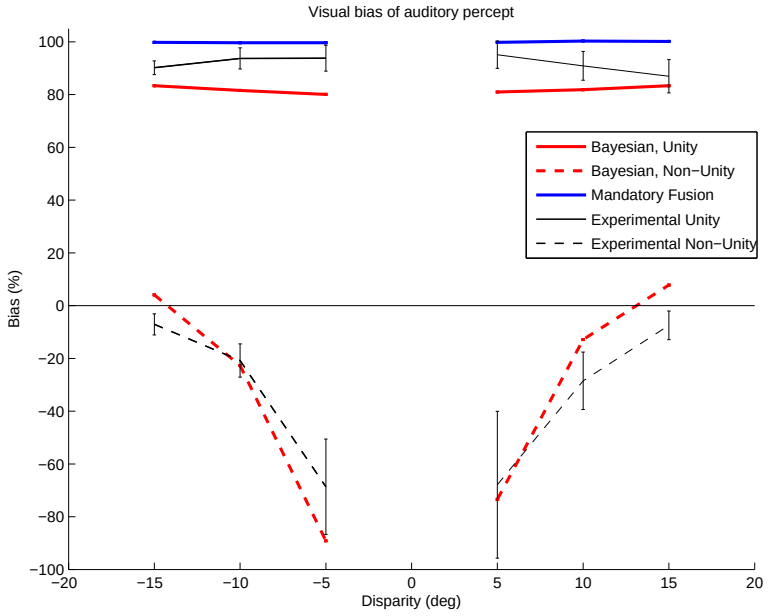


Figure 4.4: Audio-Visual gains as a function of true disparity between sources y_a and y_v. Biases observed by experiment[Wallace et al., 2004] are shown by black lines and those predicted by theory, red lines. Biases given that unity was also reported are shown by solid lines and those given that non-unity was also reported by broken lines.

Therefore, if vision totally dominated the estimate $\hat{y}_a$ of the auditory location y_a, so that on average $\hat{y}_a = y_v$, then $B = 1$. Alternately, if vision was totally ignored while estimating y_a, then on average $\hat{y}_a = y_a$ and $B = 0$.

Figure 4.4 presents the bias observed in this experiment as a function of the true disparity $(y_a - y_v)$ and whether or not the trial was perceived as unified. Wallace et al. [Wallace et al., 2004] observed strong positive bias on those trials where unity was reported (Figure 4.4, black full lines), meaning that the final auditory percept $\hat{y}_a$ moved almost completely toward the visual stimulus y_v. This is as would be expected under classical sensor fusion theory, because the more precise visual modality will dominate the estimate $\hat{y}_a$. This bias did not vary significantly with actual spatial disparity. However, a striking and unintuitive result was observed on those trials where non-unity was reported. In these cases, a zero or *negative* bias was observed (Figure 4.4, black dashed lines), where the bias was increasingly negative with smaller disparity. This meant that the auditory percept $\hat{y}_a$ moved *away* from the visual stimulus.

These surprising results are clearly reflected in the output of our model (Figure 4.4 (red lines)). The model can provide insights to help us understand these results intuitively as illustrated in Figure 4.5. The inference for U is primarily determined by

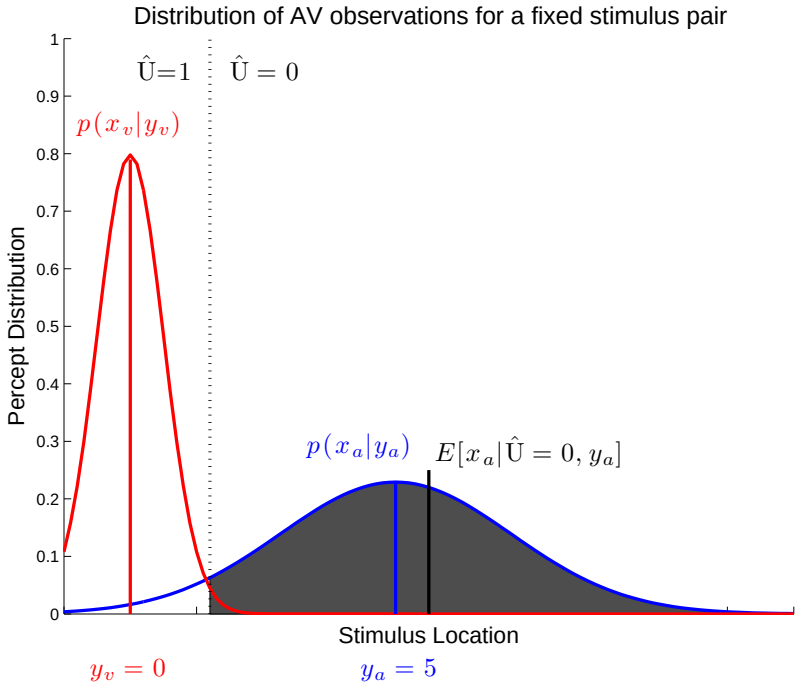


Figure 4.5: Schematic illustrating how the decision boundary for U can result in negative gains when visual y_v and auditory y_a stimuli are presented in close proximity.

the square difference between x_v and x_a (eq. (4.4)). Around observation x_v, there will therefore be a decision boundary for x_a, within which $\mathrm{U}=1$ is estimated, and outside of which $\mathrm{U}=0$ is estimated. This decision boundary effect is illustrated schematically by the dashed line in Figure 4.5.

For unity $(\mathrm{U}=1)$ to have been inferred on a given trial, the observations must have been similar, ($x_a = x_v$; either due to noise, or because actually $y_a \approx y_v$). Since this is more likely to have happened because of noise in observation x_a rather than x_v (because the visual precision is much higher), and since if $\mathrm{U}=1$ is probable, the auditory estimate $\hat{y}_a$ is pulled toward the visual observation (eq. (4.8), first term on the right), the bias will always be large and positive (Figure 4.4, upper unbroken lines).

For non-unity $(\mathrm{U}=0)$ to have been inferred on a given trial, it must have been that observation x_a was sufficiently different to x_v. If y_a and y_v were similar similar (e.g., $y_v = 0$deg, $y_a = 5$deg; Figure 4.5), then of all the audio samples $x_a \sim \mathcal{N}(y_a, p_a)$ (Figure 4.5, blue line), those which are classified as non-unified $(\mathrm{U}=0)$ will have a distribution which is *truncated* by the unification decision boundary (Figure 4.5, shaded region). The mean of this truncated distribution (Figure 4.5, shaded region) is displaced *away* from y_a – hence the surprising negative bias observed on average (Fig-

ure 4.4, lower broken lines, inner region). With similar y_a and y_v, fewer observations x_a will be estimated as U = 0 (many will lie within the decision boundary). Therefore in this region of the graph (Figure 4.4, lower broken lines, inner region), the experimental error bars are computed from fewer points, and are therefore larger. Alternately, if y_a and y_v are very different, then the size of the truncated region (and hence the extent of the bias away from y_a) is decreased – hence the tendency of the bias to climb toward zero with increasing disparity (Figure 4.4, lower broken lines, outer region).

Note that a classical sensor fusion model (Figure 4.4, blue line) cannot explain the data, as it will always exhibit nearly 100% positive bias. (In eq. (4.1), it will always infer $\hat{y}_a \approx x_v$, since $\sigma_a \gg \sigma_b$). Even in the trials where unity (U = 1) was estimated (Figure 4.4, upper unbroken lines), the Bayesian model exhibits less bias than the classical sensor fusion model, because it averages in the possibility of non-unity (eq. (4.7)).

4.2.3.2 Perception of Unity

Other observations reported in [Wallace et al., 2004] are also reflected in the inferences made by our model. Figure 4.6(a) shows the percentage of reports of unity as a function of true disparity (black lines), which are well fit by our model (red lines). This is maximum at zero disparity ($y_a = y_v$), because there, x_a will tend to be similar to x_v. However, the maximum is less than 100% because, as we have seen, noise processes occasionally result in x_a being displaced far from x_v and hence non-unity (U = 0) being inferred. At the extremes, unity reports are at minimum, but not 0%, because noise processes occasionally displace x_a toward x_v.

4.2.3.3 Localisation Error

The experimental standard deviation of the localization estimates $\hat{y}_a$ for each disparity tested is shown in Figure 4.6(b) (black lines), and again is reasonably well fit by our model (red lines). The lower standard deviation in the unified responses (Figure 4.6(b), solid lines) relative to the non-unified ones (Figure 4.6(b), broken lines) is understandable, because these are distributed primarily according to the more precise fused Gaussian statistics (eq. (4.2)). At lower spatial disparity, there is a counter intuitive increase in response standard deviation for trials perceived as non-unified. This can be understood, because for a trial with low spatial disparity ($y_a \approx y_v$) to have been perceived as non-unified, the amount of noise on the observation x_a would have had to

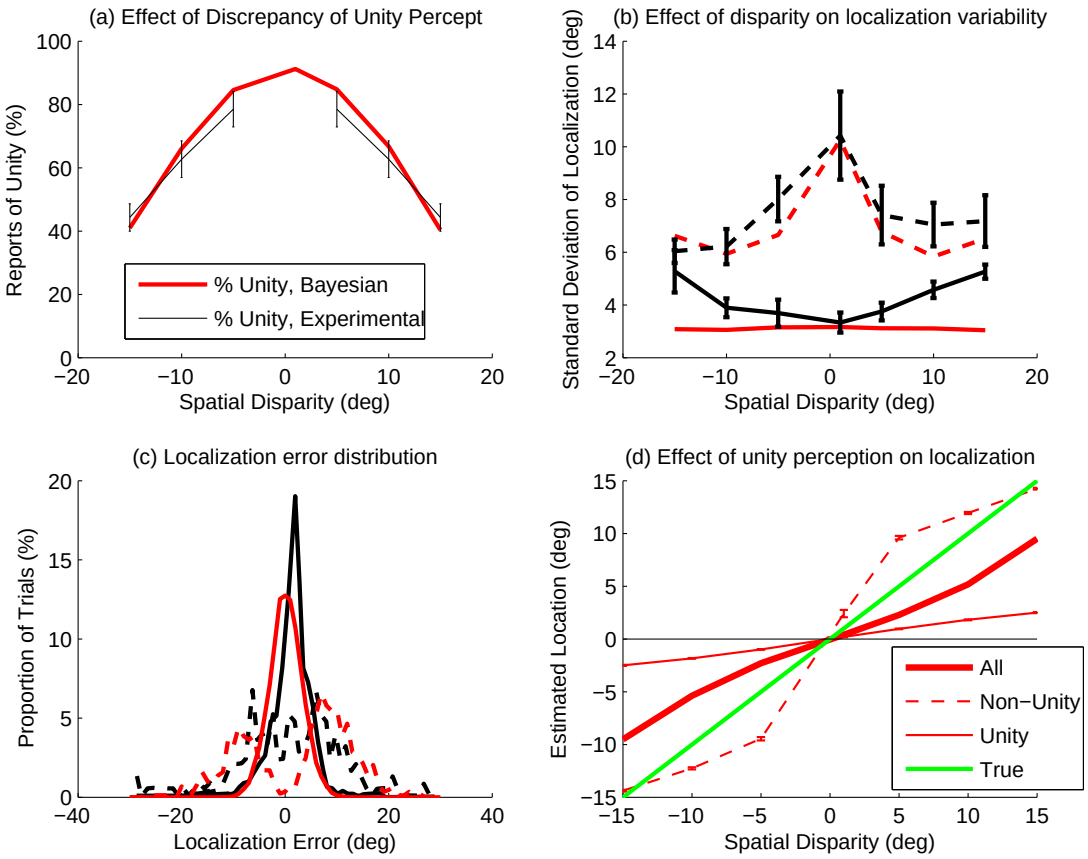


Figure 4.6: (a) Dependence of perception of unity on true discrepancy. (b) Dependence of standard deviation of localization estimates on true discrepancy. (c) Normalised histogram of localization error. (b)-(c) Experimental observations are indicated by black lines and theoretical predictions by red lines. Trials where unity was perceived are given by solid lines, and those where non-unity were perceived are given by broken lines. (d) Dependence of average location prediction on true disparity and unity percept.

be great.

The experimental distribution of errors $(\hat{y}_a - y_a)$ in the localization estimate $\hat{y}_a$ is illustrated in Figure 4.6(c), (black lines). Trials perceived as unified (solid lines) tend to have smaller localization error (because, in their location estimation, more precise visual information x_v has a stronger contribution). Trials perceived as non-unified (broken lines) have a wider distributed localization error with fewer around zero: because more precise visual information x_v has a weaker contribution. These observations are again broadly matched by our model's output (Figure 4.6(b), red lines).

4.2.3.4 New Audio-Visual Perception Predictions

Finally, one prediction we can make using this model, about a result which was not reported in [Wallace et al., 2004], is how the estimated location $\hat{y}_a$ varies as a function of disparity and reported unity (Figure 4.6(d)). Here $y_v = 0$, and the true auditory location y_a is the green axis-aligned line. All the estimates agree with this, on average, at zero spatial disparity. The trials perceived as unified (Figure 4.6(d), thin solid red line) will produce a straight line ($\hat{y}_a \approx \frac{p_a x_a + p_v x_v}{p_a + p_v}$). This line has a smaller gradient, because the higher visual precision means x_v (which is zero on average) is weighted more. The trials perceived as non-unified (Figure 4.6(d), thin broken red line) will be perceived as being displaced *away* from the true location y_a when the spatial disparity is moderate. (This is because of the asymmetrical gain effect discussed earlier from Figure 4.4.) However, as spatial disparity grows larger, the perception of the non-unified trials will gradually return to the true location y_a. In contrast to the both of the previous cases, the overall average predicted location (not sorting by unity percept) will lie between the lines of fused trails and the line of the true locations (Figure 4.6(d), thick red line).

4.2.4 Summary

Conclusions In this section, we have investigated the application of our modelling framework to understanding some recent experiments in human audio-visual perception. We modelled localization and perception of unity as the inference of latent state and model selection respectively. Recent experiments with perplexing results [Wallace et al., 2004] turn out to be well explained by our approach. We can also make novel predictions about related outputs not reported by this experiment [Wallace et al., 2004].

It is worth noting that while the classical sensor fusion equation for localisation eq. (4.1) involves a linear combination of the individual cues x_a and x_v; the structure inference solution eq. (4.8) involves a non-linear combination of cues, due to the unity posterior factors (eqs. (4.4) and (4.6)). This has interesting implications for the neurophysiological architecture of the brain which performs these computations.

It is interesting to note that broadly the same class of model (which we initially developed in Chapter 2) was able to perform useful audio-visual scene understanding (in Chapter 3) and also explain human audio-visual perception as we have just seen. This illustrates the general applicability of our approach. In Section 4.3, we will discuss the application of this approach to understanding recent experiments in visual-haptic perception.

Related Research Our theoretical approach in this section is supported by a group of very recent publications ([Kording and Tenenbaum, 2006, Beierholm et al., 2007, Kording et al., 2007, Sato et al., 2007]) from two separate labs, which independently modelled the audio-visual localization and unity experiments of Wallace et al. ([Wallace et al., 2004]). Kording et al. [Kording and Tenenbaum, 2006, Kording et al., 2007] used the same approach taken in this section – that of inferring whether or not the observations were generated from a common source (Figure. 4.3) – with qualitatively the same results and conclusions. Sato et al [Sato et al., 2007] also used the same model (Figure. 4.3). In their interpretation, however, rather than being based on the model estimate $\hat{U}$ directly, the unity report was determined indirectly: based on whether the estimates of the auditory and visual source locations were within some maximum discrepancy. This is strongly effected by the the unity estimate U, so the results and conclusions were similar. Kording et al. followed up their previous analysis of the experiments in [Wallace et al., 2004] by conducting a similar audio-visual localisation experiment of their own ([Beierholm et al., 2007, Kording et al., 2007]). In this study, they compared the structure inference approach (Figure. 4.3) with various alternatives with fixed joint priors, $p(y_a, y_v)$ – effectively marginalizing out uncertain structure U before seeing the data – and determined that the structure inference approach provided a significantly better fit to the data while requiring no additional parameters.

4.3 Visual-Haptic Oddity Detection

4.3.1 Experimental Background

A recent key experiment in the study of human sensory combination was that of Hillis, Ernst, Banks & Landy [Hillis et al., 2002], where they compared human perceptual performance using (inter-modal) multisensory cues like vision and touch as well as (intra-modal) cues like texture and disparity within vision. Their experimental paradigm involved presenting multisensory observations of three objects based on which the subject was required to discriminate the odd 'probe' object from the two 'standard' objects. Discrimination was on the basis of differences in the object height in the visual-haptic case and the object slant in the texture-disparity case.

Indeed, the oddity detection task could be performed effortlessly if the probe object was very different from the standard objects and if the probe's two multisensory observations were in agreement. However, some of the probe's sensory observations were experimentally manipulated to be discordant. i.e., no longer in agreement. Some of these discordant observations were *perceptual metamers* under the classical cue combination theory of maximum likelihood fusion. This meant that although it would be physically distinct, under this theory of cue combination, the probe object would be indistinguishable from the standard objects for the whole continuum of absolute discrepancies which formed metamers.

What Hillis et al. actually observed was a region of poor oddity detection around the point corresponding to the standard stimulus, which was elongated along the cues-discordant line. This region was significantly more extended in the intra-modal texture-slant experiment than in the inter-modal visual-haptic experiment. From this they concluded that within the senses, cues are necessarily fused (mandatory fusion), but not across the senses. The mandatory fusion conclusion was significant, because it implies that in certain multimodal perception tasks, humans have no conscious access to the unimodal observations. Their modelling and resultant conclusions, however, provide far from a complete understanding of the experimental data: The mandatory fusion theory predicts an infinite continuum of indistinguishable metameric stimuli for which oddity detection should be poor, and not merely an extended finite region of poor detection. This prediction was not supported by observations and hence, mandatory fusion fails to (even qualitatively) explain the complete data.

In this section, we present new unifying theory to model multisensory *oddity detection* and new analysis to explain the results in [Hillis et al., 2002]. We re-

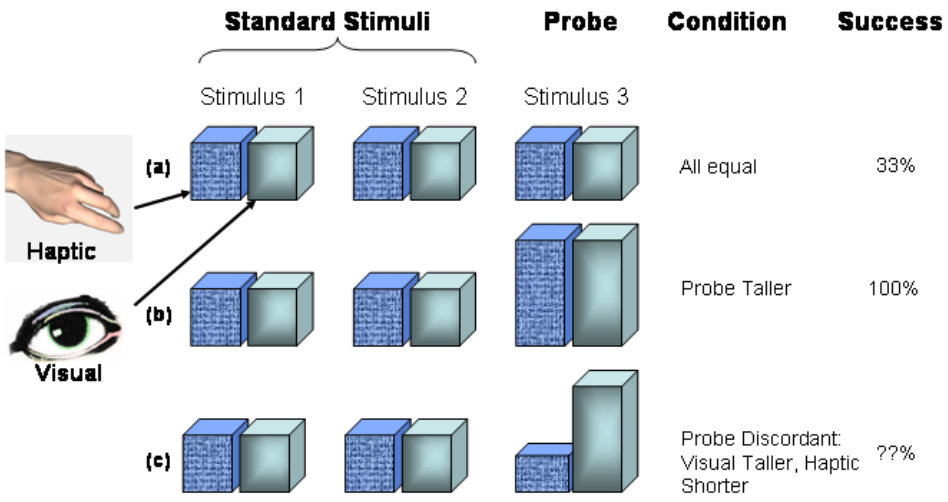


Figure 4.7: Schematic of visual-haptic height oddity detection experimental task from [Hillis et al., 2002]. Subjects must choose the odd probe stimulus based on haptic (textured bars) and visual (plain bars) observation modalities. a) Probe stimulus is the same as the standard stimuli: detection at chance level. b) Probe stimulus bigger than standard: detection is reliable. c) Haptic and visual probe modalities are discordant: detection rate will depend on cue combination strategy.

tain the general philosophy of ideal observer modelling, but apply a more accurate model of the experiment in line with the latest theory in multisensory research [Hospedales et al., 2007, Kording et al., 2007, Sato et al., 2007]. In this way, we are able to provide a quantitative yet intuitive explanation of the data, which unifies the across and within modal experiments, yet requires only one clearly interpretable free parameter. In the rest of this section, we review the experiment of interest [Hillis et al., 2002] in some detail. We introduce our new modelling framework in Section 4.3.2. Section 4.3.3 summarises our results and new predictions with some conclusions and discussion in Section 4.3.4.

4.3.1.1 Experimental Design

In the audio-visual experiments discussed so far (Section 4.2), subjects are trying to estimate a particular unknown continuous quantity y (such as height of the bar or spatial stimulus location) based on noisy observations x_i (such as visual and haptic heights or auditory and visual locations, respectively).

In the experiments of Hillis et al. [Hillis et al., 2002], which are the subject of this section, a different paradigm is used - oddity detection. Here, three stimuli are pre-

sented in two modalities[5] h and v (Figure 4.7). Two are instances of a fixed standard stimulus y_s and one is an instance of the probe stimulus y_p. The standard stimulus is always *concordant*, meaning that there is no experimental manipulation across modalities; so $y_s = y_{h,s} = y_{v,s}$. The third is a probe stimulus y_p, which is experimentally manipulated across a wide range of values so that the visual and haptic sources, $y_{v,p}$ and $y_{h,p}$, may or may not be similar to each other and to the standard y_s. The subject's task is to detect which of the three stimuli is the probe. If all the stimuli are concordant and the probe is set the same as the standard $y_s = y_p$, then we expect no better than random (33%) success rate (Figure 4.7(a)). If all the stimuli are concordant and the probe is set much greater or less than the standard $y_s \lessgtr y_p$, then we expect close to 100% success rate (Figure 4.7(b)). However, if the probe stimulus is experimentally manipulated to be *discordant* so that $y_{h,p} \neq y_{v,p}$, then the success rate expected will depend on precisely how the subjects combine their observations of $y_{h,p}$ and $y_{v,p}$ (Figure 4.7(c)). The two dimensional distribution of detection success/error rate as a function of controlled probe values $y_{h,p}, y_{v,p}$ can be measured and used to test different theories of cue combination.

For a single modality, e.g., h, the error rate distribution for detection of the probe $y_{h,p}$ can be modelled as a one dimensional Gaussian bump centred around the standard $y_{h,s}$. (If $y_{h,s} = y_{h,p}$, then detection of the odd stimulus will be at chance level. If $y_{h,p} \gg y_{h,s}$, then detection of the odd stimulus will be reliable, etc.) The shape of the two dimensional performance surface for multimodal probe stimulus detection $p(success|y_{h,p},y_{v,p})$ can be modelled as a two dimensional bump centred at (y_s, y_s). Hillis et al. [Hillis et al., 2002] compute performance *thresholds* (the equipotentials where $p(success|y_{h,p},y_{v,p}) = 66\%$) from the performance surfaces predicted by theory and those of the experimental data. The cue combination theories are evaluated by the match of their predicted thresholds to the empirical thresholds.

To parametrise models for testing, the observation precisions first need to be determined (as discussed in Section 4.1.2). Hillis et al. [Hillis et al., 2002] measure the variances of the unimodal error distributions and then, use these to predict the multimodal error distribution under mandatory fusion cue combination theory (refer eqs. (4.1) and (4.2)) — this is plausible, but as we shall see in Section 4.3.2, it is subtly different from the right thing to do.

[5]To lighten the discussion, we will refer generically to the visual-haptic $(v-h)$ modalities when discussing concepts which apply to both the visual-haptic and texture-disparity experiments.

4.3.1.2 Basic Cue Combination Theories

Hillis et al. identify a set of four basic theories (Figure 4.8) for how the brain might perform the multisensory oddity detection task, each with distinct predictions about the nature of the probe detection threshold contours (Figure 4.8, lines) around the standard stimulus (Figure 4.8, blue dot):

1. The probe stimulus might be detected based on one observation modality i only, ignoring the other entirely. This predicts a band, of width determined by the unimodal variance σ_i^2, within which the probe is too similar to the standard to be reliably detected. The band would be perpendicular to the axis of cue i and centred around the standard stimulus y_s (Figure 4.8(a), red lines).

2. The probe stimulus might be detected based on one cue and then the other, in a cascaded sequence. This predicts a rectangle about the standard y_s, within which the probe is too similar to the standard to be reliably detected. The dimensions of the rectangle are given by the intersection of the two bands from the first option, (Figure 4.8(a), red square).

3. It might compute a single fused estimate $\hat{y}_p$ based on the two observations $x_{h,p}, x_{v,p}$ (eqs. (4.1) and (4.2)) and then, discriminate purely based on this estimate. In this case, although both cues are now being used, some combinations of cues would produce a metameric probe, i.e., physically distinct but perceptually indistinguishable. Specifically, if we parametrise the probe stimuli as $y_{h,p} = y_{h,s} + \Delta y_{h,p}, y_{v,p} = y_{v,s} + \Delta y_{v,p}$, then along the line where $\Delta y_{h,p} = -\frac{\sigma_v^2}{\sigma_h^2}\Delta y_{v,p}$, the fused estimate is the same as the standard $\hat{y}_p = y_s$ and the probe would be undetectable [Hillis et al., 2002]. The band of non-detection is therefore along the cues-discordant diagonal (Figure 4.8(b), green band). The orientation and width of this band are determined by the ratio σ_v^2/σ_h^2 and $\sigma_{y|h,v}^2$, respectively. Performance along the cues-concordant diagonal is, however, improved compared to the single cue estimation cases (compare quadrants 1 and 3 in Figure 4.8(a),(b)) because, as we have seen, the combined variance is less than the individual variances ($\sigma_{y|h,v}^2 < \sigma_h^2$ and $\sigma_{y|h,v}^2 < \sigma_v^2$).

4. It might perform combined and single cue detection in sequence, giving a prediction which is the intersection of the second and third options (Figure 4.8(c), yellow area).

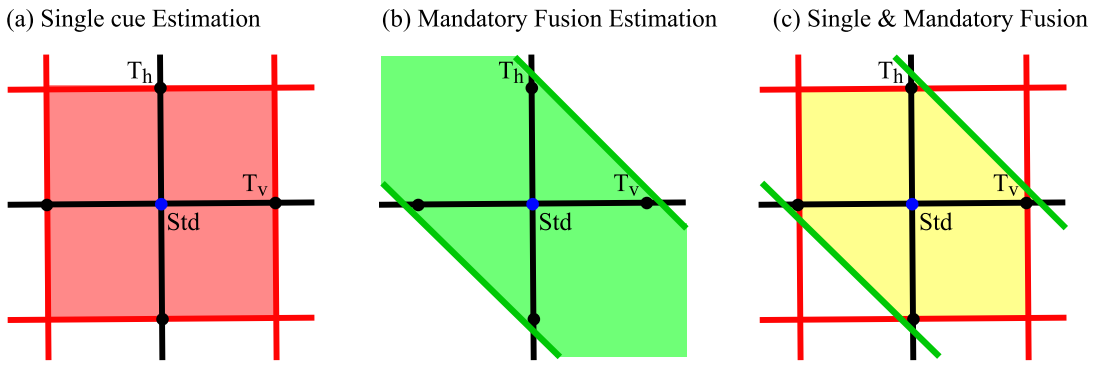


Figure 4.8: Oddity detection predictions of the basic set of cue combination models proposed by Hillis et al. [Hillis et al., 2002]. (a) Detection based on individual cues only. (b) Detection based on a single fused estimate $\hat{y}_p$. (c) Detection based on both individual cues and a single fused estimate. Shaded area indicate regions below threshold probability of correct detection. The standard stimulus y_s is indicated by a blue dot in the centre of each plot. T_v and T_h indicate unimodal visual and haptic thresholds respectively. Coloured lines indicate multimodal detection rate contours.

4.3.1.3 Results

Two variants of the experiment were performed, one for size discrimination across visual and haptic modalities (standard: $y_s = 55$mm), and one for slant discrimination using texture and stereo disparity cues within vision (standard: $y_s = 0$deg). Comparing the threshold predictions (lines) to the results observed by Hillis et al. [Hillis et al., 2002] (data points) in Figure 4.9, there are several points to note: i) In the cues concordant quadrants (1&3), the multimodal performance is increased compared to the unimodal performance, as predicted by the fusion theories (magenta points and green lines are inside the red lines in quadrants 1&3). This suggests that some cue combination is taking place, and that the first two basic theories (1,2) of independent, unimodal, detection are insufficient. ii) Particularly in the intra-modal case (Figure 4.9(b)), the observed experimental performance is significantly worse in the cues discordant quadrants (2&4) than predicted by any of the basic theories (1,2,4) which allow detection based on individual cues (magenta points are outside of the red lines in Figure 4.9(b), quadrants 2&4). In both experiments, the last basic theory (4) of sequential combined and single cue detection also fails, as performance is worse than it predicts (magenta points outside the inner bounding box of lines in Figure 4.9, quadrants 2&4).

Since the poor performance in the cues discordant quadrants 2&4 was noted to

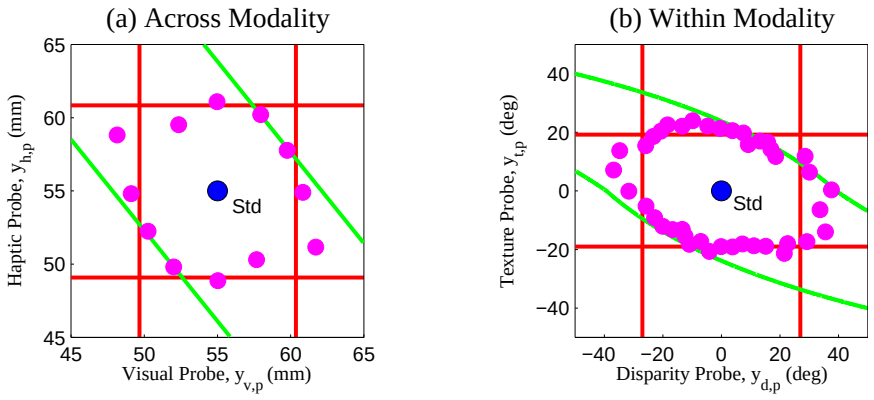


Figure 4.9: Oddity detection predictions and experimental results of Hillis et al[Hillis et al., 2002]. (a) Visual-haptic experiment. (b) Texture-disparity experiment. Red lines: Observed unimodal discrimination thresholds. Green lines: Discrimination threshold predictions assuming mandatory fusion. Magenta points: Discrimination threshold observed experimentally.

be less prominent in the inter-modal case (Figure 4.9(a)), Hillis et al. concluded that mandatory fusion applied within (Figure 4.9(b)) but not between (Figure 4.9(a)) the senses [Hillis et al., 2002]. However, even in the intra-modal case, the region of non-detection defined by the magenta points is only extended slightly away from the centre along the cues-discordant diagonal, whereas the mandatory fusion theory predicts that it should extend along an entire metameric band. The strongest conclusion that can be drawn is therefore that intra-modal perception shows a stronger tendency toward fusion than inter-modal perception.

None of the basic theories proposed (1,2,3,4) explain qualitative shape of the data well - good performance in the cues concordant quadrants 1&2 as well as a *limited* region of poor performance in the cues discordant quadrants 2&4. In particular, the classical theory of ideal observer maximum likelihood combination which Hillis et al. concluded applied in the within-modal case retains a strong *qualitative* discrepancy with the experimental results (Figure 4.9(b), green lines and points). In the next section, we will show how recent theoretical work on probabilistic models of sensor combination, which has successfully explained other related experiments, can also be applied to model multisensory oddity without the large discrepancy entailed by maximum likelihood, mandatory fusion combination.

4.3.2 Modelling Oddity

4.3.2.1 Rethinking the ideal observer model

Classical sensor fusion models (as introduced in Section 4.1.1) have been used extensively to explain human multisensory perception [Alais and Burr, 2004, Battaglia et al., 2003, Ernst and Banks, 2002, Jacobs, 1999, van Beers et al., 2002, Landy and Kojima, 2001, Gepshtein and Banks, 2003]. As we saw, the underlying motivation for this has been to test ideal observer theories of cue combination. Since these experiments are describable by the simple factored Gaussian parametric form (Figure 4.1) the optimal computations to use for inference were those described by eqs. (4.1) and (4.2).

However, the perceptual task in [Hillis et al., 2002] is not actually properly described by the standard factored Gaussian parametric form. The reason for this is that the task posed - "*Is stimulus 1, 2 or 3 the odd one out?*" - is actually no longer simply an estimation of a combined stimulus $\hat{y}_{h,v}$ or a forced choice comparison between two such combined stimuli. Such an estimation is involved in solving the task, but ultimately the task effectively asks subjects to make a *probabilistic model selection* [MacKay, 2003, Mackay, 1991] between three models[6]. This can be understood intuitively by considering the following reasoning process: *I have experienced three noisy multisensory observations. I do not know the true values of these three stimuli, but I am told that they come from two categories, standard and probe. Is it more plausible that: 1. Multisensory stimuli two and three come from one category, and stimulus one comes from a different category? Or is it more plausible that: 2. Stimuli one and three come from one category, and stimulus two comes from a different category? Or: 3. Stimuli one and two come from one category, and stimulus three comes from a different category?*

With this in mind, to take a Bayesian ideal observer point of view on this experiment, we clearly need a slightly more sophisticated model selection approach than the simple factored sensory fusion approach of Section 4.1 and Figure 4.1. This should *integrate over the distribution of unknown stimulus values* y_s and y_p (since subjects are not directly asked about these) in determining the most plausible model (assignment of oddity).

A generative model Bayesian network formalisation of the oddity detection task for

[6] It can also be understood as finding the most likely assignment of points in a clustering task. Specifically, consider mixture of Gaussian clustering of three two-dimensional points into two clusters with unknown means.

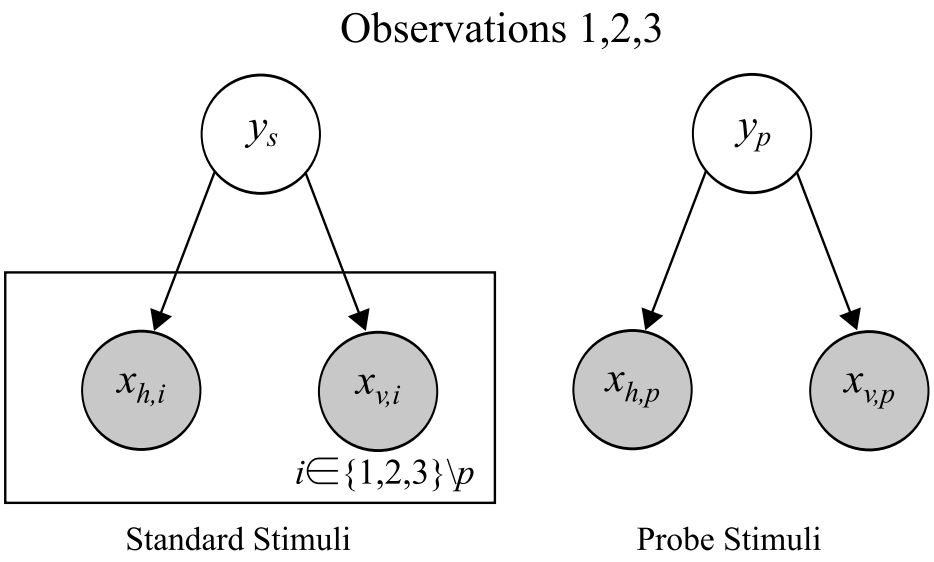


Figure 4.10: Model for oddity detection by model selection. There are three possible models, indexed by p, corresponding to each possible assignment of oddity. To compute the stimulus most likely to be odd, compute the evidence for each model $p(\{x_{h,i}, x_{v,i}\}_{i=1,2,3}|p)$. Standard and probe stimulus values y_s, y_p are not directly requested of the subjects, and are only computed indirectly in the process of evaluating the model likelihoods.

the three multisensory observations $\{x_{h,i}, x_{v,i}\}_{i=1}^3$ is shown in Figure 4.10, where the task is to determine which observation is the probe. The graph on the right indicates that the probe visual-haptic observations are related via their common parent, the latent probe stimulus of value y_p. The graph on the left indicates that the four observations composing the other two standard stimuli are all related to the standard stimulus value y_s. The three different instantiations of this model are given by the different probe hypotheses $p = 1,2,3$ which separate the standard and probe stimuli into different clusters (via the set difference operator, $'\setminus'$ in our notation). For example, $p = 3$ would mean that observations $\{x_{h,1}, x_{v,1}, x_{h,2}, x_{v,2}\}$ (Figure 4.10, left) should be similar to each other (all being drawn from the standard y_s) and potentially dissimilar to observations $\{x_{h,3}, x_{v,3}\}$ (Figure 4.10, right), which were generated independently from y_p. With uniform prior belief about which stimulus p is the probe, the ideal Bayesian observer would compute the evidence $p(\{x_{h,i}, x_{v,i}\}_{i=1}^3|p,\theta)$ for each of the three models $p = 1,2,3$ as:

$$p(\{x_{h,i}, x_{v,i}\}_{i=1}^3|p,\theta) \quad = \quad \int_{y_s} p(\{x_{h,i}, x_{v,i}\}_{i\in\{1,2,3\}\setminus p}, y_s|p,\theta)$$

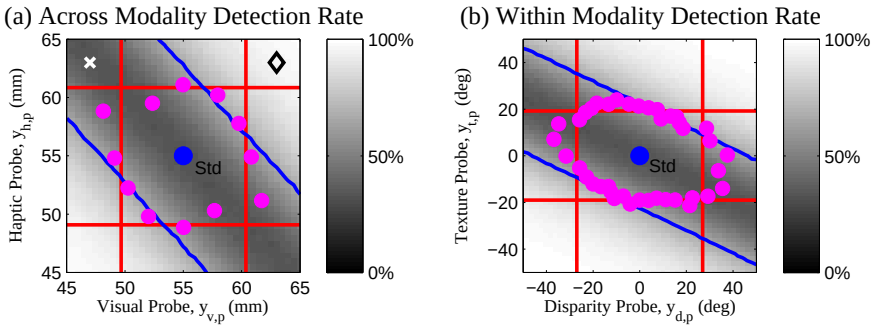


Figure 4.11: Model oddity detection performance as a function of probe value (grey-scale) with 66% contours (lines) for comparison with human performance (dots). This model still predicts an infinite region of non-detection along the cues-discordant diagonal. (a) Across modality visual-haptic experiment. (b) Within modality texture-disparity experiment.

$$\begin{aligned} & \cdot \int_{y_p} p(x_{h,p}, x_{v,p}, y_p | p, \theta), \qquad (4.10) \\ = \; & p_s(\{x_{h,i}, x_{v,i}\}_{i \in \{1,2,3\} \setminus p} | p, \theta) p_p(x_{h,p}, x_{v,p} | p, \theta). \end{aligned}$$

and report the model with the highest likelihood $\hat{p} = \text{argmax}_p p(\{x_{h,i}, x_{v,i}\}_{i=1}^3 | p, \theta)$. Here, θ summarises all the fixed model parameters, e.g., the observation variances σ_h^2 and σ_v^2. This model evidence evaluation procedure integrates over the specific latent stimuli values y_s and y_p, as subjects are not directly asked about them. In the event that all distributions involved are Gaussian, eq. (4.10) is simple to evaluate (see Appendix A.2.2, for detailed parametric form and derivation).

This model (Figure 4.10, eq. (4.10)) predicts probe detection only outside of the cues-discordant diagonal (Figure 4.11(a),(b), lines), which is qualitatively similar to the simple factored fusion model (Figure 4.8(b)) and still does not match the data (Figure 4.11(a),(b), points).

Some intuition about how this works can be gained by considering the form of the entire normalised data distribution $p(\{x_{h,i}, x_{v,i}\}_{i=1,2,3} | p, \theta)$ for each model p [MacKay, 2003], which in this case factorizes into a standard and probe component (eq. (4.10)). For example, the model $p = 3$, predicts that the probability mass of the distribution of observations $\{x_{h,1}, x_{h,2}, x_{v,1}, x_{v,2}\}$ should lie around a four dimensional line through the standard (where $x_{h,1} = x_{h,2} = x_{v,1} = x_{v,2}$) while the distribution of probe observations $\{x_{h,3}, x_{v,3}\}$ should lie around the line where $x_{h,3} = x_{v,3}$ in two dimensional space. Assuming, without loss of generality, that the true model is $p = 3$, then observations at the point indicated by the diamond in Figure 4.11(a) will be correctly

classified: The correct model $p = 3$ will have high likelihood, as the first four observations will be very similar and lie within the standard probability mass, and the two probe observations will be similar to each other and lie within the probe probability mass. An incorrect model, e.g., $p = 1$, will have low likelihood because the observation $\{x_{h,2}, x_{h,3}, x_{v,2}, x_{v,3}\}$ are not at all similar, and so do not lie within the standard probability mass.

Consider instead the point indicated by the cross in Figure 4.11(a). Here, under the hypothesis that $p = 3$, while the standard observations do lie within the standard probability mass, the discordant probe observations do not lie within the probe probability mass (which was around the line where $x_{h,3} = x_{v,3}$), so this hypothesis is unlikely. However, the other hypotheses are also unlikely. For example, consider the alternative $p = 1$, then although $\{x_{h,1}, x_{v,1}\}$ does lie within the probe mass, the remaining observations $\{x_{h,2}x_{h,3}, x_{v,2}, x_{v,3}\}$ have discordant components and now no longer lie within the standard mass. Therefore no one model is clearly the most probable, and detection is unreliable.

We should not expect the ideal observer model to explain the empirical data just yet, however, as there is one final aspect to the task which has not yet been modelled. This is the structure uncertainty, which we discuss next.

4.3.2.2 Structure Inference

All of the oddity detection models discussed so far (Figures 4.1 and 4.10) have assumed a fixed structure. Recent multisensory perception experiments [Hillis et al., 2002, Hairston et al., 2003, Wallace et al., 2004, Shams et al., 2000, Shams et al., 2005, Roach et al., 2006] have, however, presented subjects with what is essentially a variable causal structure. It is therefore unsurprising that the simple fixed structure ideal observer models have failed to explain the results. As we saw in Section 4.2, and as was argued recently in the literature [Kording et al., 2007, Sato et al., 2007, Hospedales et al., 2007], the results of these experiments can be explained by extending the underlying Bayesian model appropriately.

The new variable introduced in these recent experiments is uncertainty in whether a given pair of observations are actually related or not. In Section 4.2 we considered the experiments reported in [Hairston et al., 2003, Wallace et al., 2004]. Here, subjects were presented with stimuli from a range of audio and visual stimulus positions; so some were concordant and others were not. They were asked to point out where they thought the audio stimulus came from and whether they thought the visual stim-

ulus co-occurred with the audio stimulus. When the audio and visual stimuli were similar, a unified percept was reported and the reported position was approximately the weighted average of the stimulus as we might expect from maximum likelihood integration [Kording et al., 2007, Sato et al., 2007]. When the stimuli were very discrepant, they were reported to be non-unified, and the position report showed no or negative interaction. The extra uncertainty here is whether the multisensory stimuli did indeed come from the same source or not. This is equivalent to posing uncertain causal structure in the probabilistic model for the ideal observer. We introduced the approach needed to solve this type of problem in multisensory perception as *structure inference* [Hospedales et al., 2007]. Kording et al. [Kording et al., 2007] carried out a detailed analysis of these experiments [Hairston et al., 2003, Wallace et al., 2004] and showed that the structure inference approach was necessary to explain the results, but termed the procedure *causal inference*.

Structure Inference in Oddity Detection Returning to the oddity experiment of interest, the region of the probe stimulus space not explained by current models is that in which Hillis et al. [Hillis et al., 2002] have manipulated the multisensory probe observations such that they have implausibly large cross-modal discrepancy. In doing so, they have introduced variability which the models so far (Figures 4.1 and 4.10) cannot represent, so it is unsurprising that they do not predict the data well (Figures 4.8 and 4.11). Specifically, in the regions of data discrepancy, probe stimuli $y_{h,p}$ and $y_{v,p}$ are discordant enough that even the model in Figure 4.10 (which represents the probe stimulus using only one variable for both modalities y_p) is no longer a plausible explanation of the observations.

The relevance of this to the experimental results becomes evident when we note that only the probe stimuli can have discordant (inconsistent) multimodal observations. Therefore, the subjects could potentially detect the probe on the discordant-cues axis (on which neither of the models so far can detect the probe) if they can infer this *change in structure* – a potential explanation for the exact source of discrepancy identified earlier between the observed results and our model so far. Indeed in their post experimental analysis, Hillis et al. [Hillis et al., 2002] noted that, "*Sometimes [the subjects] used a difference in perceived size, but frequently they noticed the conflict between the visually and haptically specified sizes and used the perceived conflict to make the oddity discrimination.*". Although unlike [Hairston et al., 2003, Wallace et al., 2004], Hillis et al. did not systematically ask sub-

jects for their perception of multisensory unity or not for each stimulus, this comment strongly suggests that the subjects in [Hillis et al., 2002] did infer and use the information about the unusual structure in their task (as they have in other related experiments [Kording et al., 2007, Hairston et al., 2003, Wallace et al., 2004]). Next, we formalize how to model the structure uncertainty in this experiment.

Modelling Structure Inference Our model selection interpretation of the oddity detection problem (Figure 4.10), can easily be updated to take into account the potential dis-association of the two probe stimulus modalities as shown in Figure 4.12. (Note that the original simple factored model (Figure 4.1, [Hillis et al., 2002]) cannot be updated in this way.) Here, the Bernoulli association variable C has been introduced to represent the uncertain structure; whether the multisensory probe observations have a common source or not. This unavoidably introduces the free parameter π_c in the prior for C, $p(C) = \pi_c^C(1-\pi_c)^{(1-C)}$. If we were certain a-priori of common causation ($\pi_C = 1$), we then have the special case of the model from Figure 4.10. If $0 < \pi_c < 1$, then while computing the evidence for each model $p(\{x_{h,i}, x_{v,i}\}_{i=1}^3|p)$, we integrate over the causal structure C, e.g., whether we are feeling and seeing the same thing or not. The exact value of π_c used will depend on particular combination of senses or cues being used and the particular context and task (and it may vary between people, as do σ_v^2, σ_h^2 etc). Under the hypotheses of common causal structure $C = 1$, we assume that the two observations $x_{h,p}, x_{v,p}$ were produced from a single latent variable y_p, while under the alternate hypothesis $C = 0$, we assume separate sources $y_{h,p}$ and $y_{v,p}$ were responsible for each.

$$p(\{x_{h,i}, x_{v,i}\}_{i=1}^3|p,\theta) = \sum_C \int p(\{x_{h,i}, x_{v,i}\}_{i=1}^3, y_s, y_p, y_{h,p}, y_{v,p}, C, |p,\theta) dy_s dy_p dy_{h,p} dy_{v,p}. \quad (4.11)$$

To evaluate the likelihood of each stimulus being the probe, the ideal Bayesian observer would compute the model likelihoods $p(\{x_{h,i}, x_{v,i}\}_{i=1}^3|p,\theta)$ in eq. (4.11). This is again simple to compute if all the stimulus distributions are Gaussian, requiring only numerical integration of the binary causal structure variable, C. The specific parametric solution used is given in the Appendix A.2.3.

Model Parameters Now we discuss how we set the four parameters of this model: The noise level on each modality, e.g., σ_v^2, σ_h^2, the prior belief about the distribution

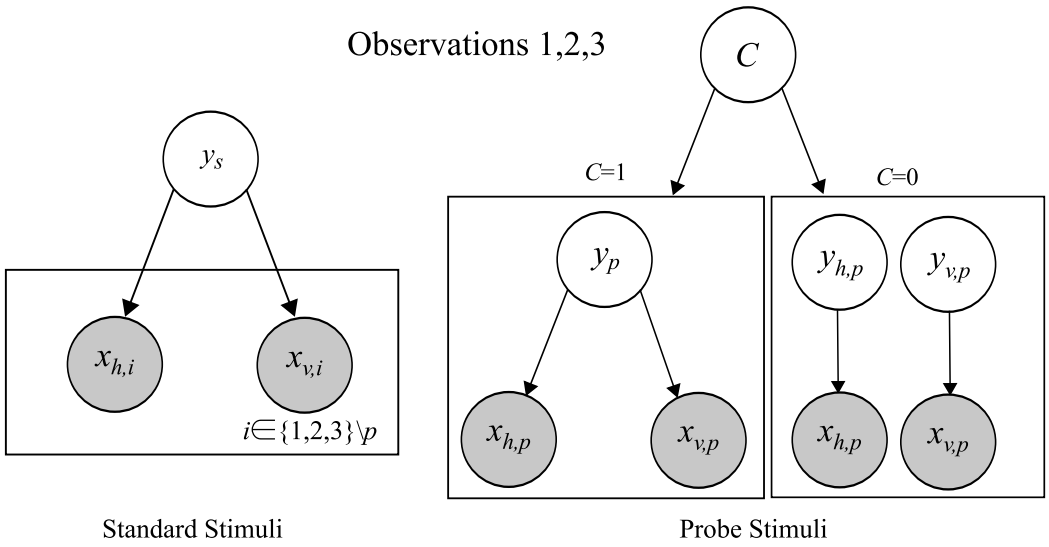


Figure 4.12: Graphical model for oddity detection via structure inference. Again the three possible assignments of oddity correspond to three possible models indexed by $p = 1,2,3$. The uncertainty about common causal structure in of the probe stimulus is now represented by C which is computed in the process of evaluating the likelihood of each model p.

over bar heights y, and the prior probability of fusion π_c.

We cannot use any of the standard parametrisation methods discussed specifically in Section 4.1.2 in our interpretation of the experiment as a discrete three way model selection problem. However, we can still set the unimodal variances σ_v^2, σ_h^2 in an analogous way - by matching the simulated unimodal experiment to the unimodal experimental results (Figure 4.9, red lines). Specifically, we take the model of eq. (4.11), Figure 4.12 and consider only one modality at a time. (Without using the extra structure variable as this is only relevant for multimodal observations.) For any given setting of σ_i^2, we can simulate the whole unimodal experiment and measure the 66% performance threshold. So, we simply perform a one dimensional search to find the value of σ_i^2 which produces the threshold most closely matching the unimodal experimental data (Figure 4.9, red lines).

For a Bayesian model, we are unavoidably required to specify some prior belief about the latent stimulus sizes y, and it is mathematically convenient for these to also use a Gaussian parametric form $p(y) = \mathcal{N}(y; \mu_y, \sigma_y^2)$. We use the same distribution for all the latent ys. We assume subjects have correctly estimated the true mean μ_y of the latent distributions which is the standard stimulus – 55mm in the intra-modal experiment and 0deg in the inter-modal experiment. The variance of the subjects' prior

belief σ_y^2 is slightly harder to set appropriately. We set the prior for each experiment relatively uninformatively ($\sigma_y = 20$mm and $\sigma_y = 100$deg) for all subjects to ensure the whole state space investigated by the experiment was plausible under the prior distribution. Subsequent analysis showed that, unlike for σ_h^2, σ_v^2, the results are in any case highly insensitive to the specific value of σ_y^2.

Finally, we expect the prior probability of fusion π_c to be somewhat dependent on the subject and the within versus across-modal experimental conditions. In the subsequent analysis, we coarsely fit π_c for each subject and experimental condition to minimise the mean square error between the predicted and experimental contours (refer Section 4.3.3.1 for details).

With the complete and parametrised model of the experiment structure, we now expect that an ideal observer using this model for inference should explain the data. In the next section, we compare the predictions of our 'variable structure' oddity detection model (Figure 4.12, eq. (4.11)) to the experimental data.

4.3.3 Results

To evaluate our multisensory oddity detection model as developed in Section 4.3.2, we compute the success rate distribution produced by our model when detecting the probe, $\hat{p} = \text{argmax}_p p(p|y_s, y_{h,p}, y_{v,p})$, as a function of the probe values $y_{v,p}$ and $y_{h,p}$. We can then compare the 66% performance thresholds of the model's success rate distribution $p_m(\hat{p}_{\text{correct}}|y_s, y_{h,p}, y_{v,p})$ against the human success rate distribution $p_e(\hat{p}_{\text{correct}}|y_s, y_{h,p}, y_{v,p})$ as measured in [Hillis et al., 2002] (Figure 4.9, dots).

A subtle but important point to note for doing this correctly is that on any particular trial, while the experimenter controls the stimuli $y_s, y_{h,p}, y_{v,p}$, the human subject uses the noisy observations $\{x_{h,i}, x_{v,i}\}_{i=1,2,3}$ as input for their computation. It is, therefore, insufficient to simply control $\{x_{h,i}, x_{v,i}\}_{i=1,2,3}$ and compute the model's response $p_e(p|\{x_{h,i}, x_{v,i}\}_{i=1,2,3})$ as this is not what is being reported by the experiment in [Hillis et al., 2002]. To produce truly comparable results to that of the human experiment, we need to compute the model output as a function of the *pre-noise* input (y_s, y_p) as is controlled in the human experiments. We, therefore, integrate over the actual noisy observations $\{x_{h,i}, x_{v,i}\}_{i=1,2,3}$ as follows:

$$\begin{aligned} p_m(p|y_s, y_{h,p}, y_{v,p}) &= \int p(p|\{x_{h,i}, x_{v,i}\}_{i=1,2,3}) \\ &\quad \cdot p(\{x_{h,i}, x_{v,i}\}_{i=1,2,3}|y_s, y_{h,p}, y_{v,p}) dx_{h,1} dx_{h,2} dx_{h,3} dx_{v,1} dx_{v,2} dx_{v,3}, \end{aligned}$$

$$= \int p(p|\{x_{h,i},x_{v,i}\}_{i=1,2,3})p(x_{h,p}|y_{h,p})p(x_{v,p}|y_{v,p})dx_{h,p}dx_{v,p} \cdot \prod_{j=\{1,2,3\}\setminus p} p(x_{h,j}|y_s)p(x_{v,j}|y_s)dx_{h,j}dx_{v,j}. \quad (4.12)$$

We approximate this by sampling 50,000 noisy observations $\{x_{h,i},x_{v,i}\}_{i=1,2,3}$ for every probe condition $y_s,y_{h,p},y_{v,p}$ and averaging over the response of the model to each sample. The importance of correctly simulating the noise processes in psychophysics models was recently discussed in the analysis of a related experiment [Kording et al., 2007]. The measured $p_e(\hat{p}_{\text{correct}}|y_s,y_{h,p},y_{v,p})$ for human subjects can now be correctly and directly compared to the success rate of the model $p_m(\hat{p}_{\text{correct}}|y_s,y_{h,p},y_{v,p})$.

4.3.3.1 Bayesian Multisensory Oddity Detection Results

Detection Threshold Contours Figures 4.13(a) and (b) illustrate the across and within modality results, respectively. The experimental data (dots) are shown along with the global performance of the model across the whole input space (grey-scale background, with white indicating 100% success) and the 66% performance contour (blue lines). The human experimental measurements broadly define a region of non-detection centred about the standard and slanted along the cues discordant line and stretched slightly outside the bounds of the inner unimodal threshold rectangle. The extent of the non-detection region along this line is increased somewhat in the within modality case as compared to the across modality case [Hillis et al., 2002].

As discussed in Section 4.3.1.2, none of the simple models – single cue based estimation (Figure 4.8(a), red lines), mandatory fusion (Figure 4.8(b), green lines) or combination thereof – explain these particular observations. Moreover, the classical maximum likelihood mandatory fusion theory makes the qualitative error of predicting infinite bands of indiscriminability (Figure 4.8, green lines). In contrast, our Bayesian model provides an accurate quantitative fit to the data (Figure 4.13, blue lines).

To quantify this, we followed [Hillis et al., 2002] in computing the distance from the standard to each experimental threshold point and the closest predicted threshold along the vector to that point (Figure 4.13, points and lines). We could then compare the root mean square error (RMSE) between the experimental threshold distance and the threshold distance predicted by the various models. The qualitative discrepancy between the data and the solely unimodal or solely mandatory fusion models is clearly highlighted by this measure: Since for many experimental data points there are no predicted thresholds on that vector, these models have infinite error. The two

remaining simple models were based on sequentially testing each unimodal cue independently (Figure 4.13(a), red rectangle) and sequentially testing the fused estimate followed by each unimodal cue independently (Figure 4.13(c), yellow region). We therefore compared our Bayesian model against the sequential unimodal and sequential fusion models which had RMSE of 0.8mm, 0.9mm and 1.1mm respectively in the across-modality experiment and RMSE of 2.6deg, 3.9deg and 5.0deg respectively in the within-modality experiment. Our Bayesian ideal observer model therefore provides the best quantitative match to the data as well as the only explanation of the data's specific qualitative form: good performance in quadrants 1&3 as well as a *limited* region of poor performance in quadrants 2&4.

To produce these contours, we fit the prior probability of fusion p_c to the data, so as to minimise the contour error, determining $\pi_c^{across} = 0.995$ and $\pi_c^{within} = 0.999$, which are plausible values. It is very unlikely that different visual cues at the same retinal location are due to different objects, hence the stronger prior for fusion within vision. Seeing and manipulating different objects simultaneously is somewhat more common, so the weaker prior for fusion in the across modality case is expected.

Perception of Fusion To gain some intuition into this, we can again consider the normalised distribution of the data eq. (4.11) under each model here as compared to the fixed structure case discussed in Section 4.3.2.1, eq. (4.10). Now, after marginalising over C, the probability mass in the probe part of this distribution is a mixture, spread both around the line $x_{h,p} = x_{v,p}$ as before ($C = 1$) and also more uniformly over the space ($C = 0$). Therefore, multisensory observations involving sufficiently discordant points are relative plausible under the probe distribution, allowing points in quadrant 2&4 to be correctly classified; which was not possible in the example described in Section 4.3.2.1.

To understand clearly how the Bayesian model works, we can also consider its marginal inference for the fusion (common multisensory source) of the probe $p_m(C|y_s, y_{h,p}, y_{v,p})$, shown in Figure 4.13(c),(d). This corresponds to the human answer to the question "*Do you think your visual and haptic observations are caused by the same object, or have they become discordant?*". This question was unfortunately not asked systematically in [Hillis et al., 2002], but the subjects' self reporting of a detection strategy based on discordant cues is in line with the strategy that falls out of inference with our model.

Along the cues concordant line, the model has sensibly inferred fusion (Fig-

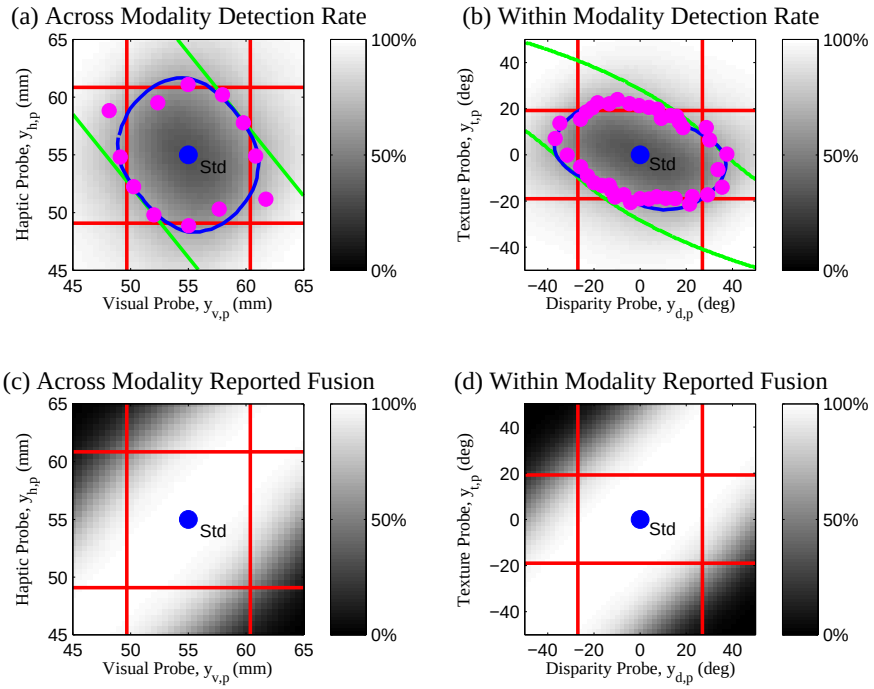


Figure 4.13: (a,b) Oddity detection rate predictions for an ideal Bayesian observer (grey-scale background) using a variable structure model; Oddity detection contours of the model (blue lines) and human (magenta points) are overlaid with the Hills et al. [Hillis et al., 2002] model prediction (green lines); Chance=33%. (c,d) Fusion report rates for ideal observer using variable structure model. Chance=50%. Across modality conditions are reported in (a,c) and within modality conditions are reported in (b,d).

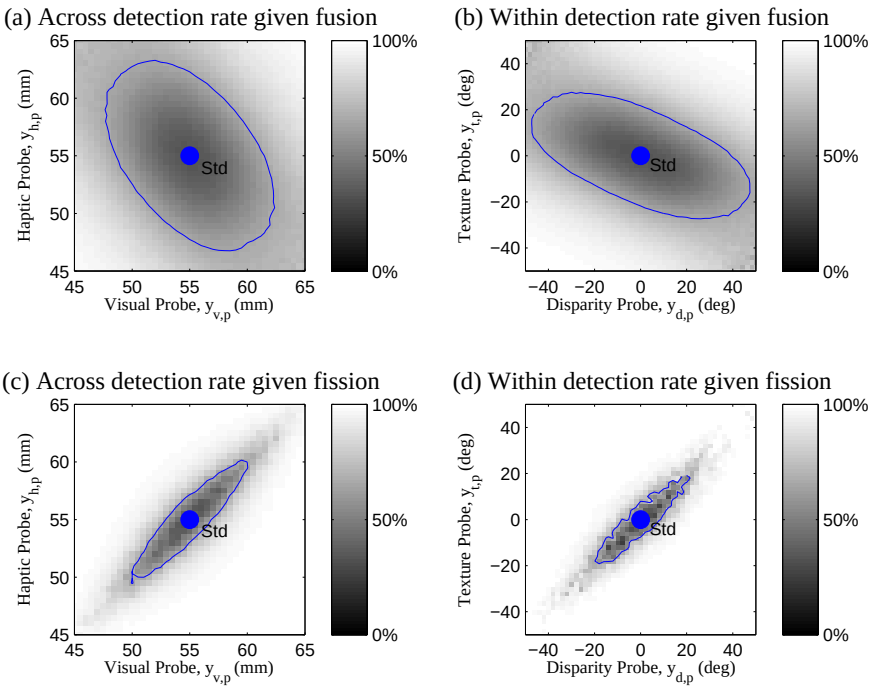


Figure 4.14: New predictions by the ideal Bayesian observer using the variable structure model. (a,b) Detection rate for trials where fusion was reported (Chance = 33%). (c,d) Detection rate for trials where fission was reported (Chance = 33%). Across-modality condition in (a,c), within modality condition in (b,d). Blue lines indicate contours of detection threshold (66%).

ure 4.13(c),(d), quadrants 1&3). In these regions, the model can effectively detect the probe (Figure 4.13(a),(b), quadrants 1&3), and the fused probe estimate $\hat{y}_p$ is different to the standard probe estimate $\hat{y}_s$.

On the other hand, considering trials moving away from the standard along the cues discordant line, the model eventually infers fission (Figure 4.13(c),(d), quadrants 2&4). The model infers the probe stimuli correctly in these regions (Figure 4.13(a),(b), quadrants 2&4) where the mandatory fusion models cannot (Figure 4.13(a),(b), quadrants 2&4, green lines) because the probe and standard estimates would be the same $\hat{y}_p = \hat{y}_s$. The strength of discrepancy between the cues required before the fission is inferred depends on the variance of the observations (σ_h^2 and σ_v^2) and the strength of the fusion prior π_c, which will vary depending on the particular task and combination of modalities. A total of nine experiments were reported in [Hillis et al., 2002]. Appendix A.2.1 includes the resultant fits of our model to the remaining experiments along with the comparative error analysis (RMSE) to the other models.

4.3.3.2 New Oddity Detection Predictions

The internal working of the Bayesian model developed here provide new directly testable predictions about human behaviour in this task. If the participants were also asked for their percept of fusion/fission as well as their oddity estimate (e.g., as in the audio-visual experiments [Hairston et al., 2003, Wallace et al., 2004]), then the model makes some specific and surprising predictions for oddity detection rate as a function of whether a given trial was also perceived as fused or not. These are illustrated in Figure 4.14.

- Although overall performance for detecting probes away from the standard was good (Figure 4.13(a),(b), all quadrants), for those trials where fusion was specifically reported, the discrimination will be more reliable *off* the cues-discordant axis (Figure 4.14(a),(b)). Explicitly, see the increased extent of the detection threshold contour along the cues discordant axis in Figure 4.14(a),(b) compared to Figure 4.13(a),(b).

- More strikingly, for those trials where fission was reported, the discrimination will only be reliable *off* the cues-concordant axis (Figure 4.14(c),(d)). This is the opposite effect to that of trials overall (Figure 4.13(a),(b)) and fused trials (Figure 4.14(a),(b)). To gain some intuition about this, consider that for a cues-concordant trial to have been inferred as fission, there must have been unusually large noise separating the observations $x_{h,i}$ and $x_{v,i}$ composing the particular multimodal stimulus i which were inferred to be the probe. However, this event would be just as *un*likely in happening to a pair of the true standard observations (causing wrong probe identification) as it would be in happening to the pair of true probe observations. Hence, probe detection under these circumstances would be unreliable.

4.3.4 Summary

In this section, we have developed a Bayesian ideal observer model for multisensory oddity detection and used it to re-examine the experiments of Hillis, Ernst, Banks & Landy [Hillis et al., 2002]. In [Hillis et al., 2002], the standard maximum likelihood ideal observer approach seemed to fail with drastic qualitative discrepancy compared to the empirical results; however, this was due to an subtly inappropriate model. With the complete model of the experimental task developed here, the Bayesian ideal observer

approach provides an accurate quantitative explanation of the data with only one free parameter π_c which represents a clearly interpretable quantity: prior probability of common causation. Optimization intuitively sets it to be greater in the within modality case than the across-modality case.

Two novel steps were required to correctly model this problem. The first was the understanding of the problem as a model selection task related to clustering. The unknown bar size or surface slant is of key consequence for the oddity detection, but not directly reported and should therefore be modelled, but integrated over, by a Bayesian observer. Our interpretation of the problem is also satisfying in that all the variables in the model represent concrete physical quantities. (E.g., haptically observed bar height $x_{h,i}$ for each object i, unknown discrete index p of the odd object.) This is unlike the original analysis [Hillis et al., 2002], which attempted to model the detection rate contours directly without, for example, a notion of which particular object p was odd: a quantity which the brain is clearly computing, as it is the goal of the task. Moreover, within the field of perceptual modelling, we are interested in possible computational mechanisms of inference – in this case for p – which we propose here, but which were not proposed originally [Hillis et al., 2002].

The second novel step required was the use of a model with variable structure to appropriately reflect the subject's uncertainty in the causal structure C of their observations due to the experimental manipulation. This structure inference approach [Hospedales et al., 2007] has recently been used to understand other similarly perplexing experimental results in human audio-visual multisensory perception [Kording et al., 2007, Hairston et al., 2003, Wallace et al., 2004, Shams et al., 2000, Shams et al., 2005].

In summary, the standard maximum likelihood integration approach to sensor fusion has dramatically failed to explain the experimental data in [Hillis et al., 2002]. This data can now be understood as result of the perceptual system, as a Bayesian ideal observer, computing the most likely probabilistic model for noisy data under uncertain causal structure. This theory provides an accurate and intuitive explanation of the data and, via the parameter π_c, unifies the within and across-modal scenarios.

4.3.4.1 Related Research

The framework proposed may seem more complicated than the simple factored cue combination approach ([Hillis et al., 2002], Figure 4.1). However, this is necessary and appropriate, because the actual experimental task of oddity detection under causal

structure uncertainty is more complicated than the classical task of stimulus estimation by cue combination. E.g., we represent, and propose a mechanism for inference of p which is not included in [Hillis et al., 2002] (Figure 4.1). Our approach is parsimonious, in that within the research theme of investigating the extent to which human perception is Bayesian optimal [Ernst and Bulthoff, 2004, Knill and Richards, 1996], models should use the same generative process as the perceptual experiment. By modelling the three sets of stimuli, including the selection of a probe stimulus and potential disassociation within that stimulus, we have done just this – and provided the best explanation of the data. Finally, despite any apparent complexity, the new model introduces only one new parameter compared to the models in [Hillis et al., 2002].

The theory and practice for modelling uncertain causal structure in inference tasks has a more extensive history in other fields. In artificial intelligence, the theory goes back to Bayesian multinets [Geiger and Heckerman, 1996], and is applied today, for example, in building artificial systems to explicitly understand audio-visual correlations [Hospedales and Vijayakumar, 2008] for multisensory speaker detection. In radar tracking, this problem is known as data association [Bar-Shalom and Tse, 1975], and its solutions are used to sort out multiple radar detections of uncertain causal relation to multiple aeroplanes into consistent and accurate estimates of the aircraft locations.

A variety of recent multisensory cue combination studies have reported cue fusion when the cues are similar and fission when the cues are dissimilar [Hairston et al., 2003, Wallace et al., 2004, Shams et al., 2000, Shams et al., 2005, Roach et al., 2006]. Recently, some authors have tried to understand these type of effects as being the result of a correlated joint prior over the multisensory sources like y_h and y_v. This may be Gaussian in their difference [Ernst, 2005, Bresciani et al., 2006], reflecting a prior belief that they should be similar. This prior is insufficient, as it can not explain complete segregation (complete non-interaction of the observations) observed in many experiments, because the jointly Gaussian prior always attracts the estimates of the stimuli together. Alternately, the joint prior may take the special form of a Gaussian-uniform sum [Roach et al., 2006], to reflect the fact that the observations are sometimes correlated and sometimes not. This is related to our model in that if we chose not to represent structure uncertainty C, and simplified our generative model by $p(y_{h,p}, y_{v,p}|\theta) = \sum_C p(y_{h,p}, y_{v,p}|C, \theta)p(C|\theta)$, then the joint probability of the visual and haptic stimuli would have approximately a Gaussian-uniform sum form. Inference of the probe stimulus values $y_{h,p}, y_{v,p}$ is this case would tend to be fused if the observations

$x_{h,p}, x_{v,p}$ were similar and independent if the observations were dissimilar. However, this would be unsatisfactory as the experiment would now not be as accurately represented by the model. Moreover, the model would then not explicitly represent the structure C, which subjects do infer explicitly as reported in [Hillis et al., 2002] and other related experiments [Hairston et al., 2003, Wallace et al., 2004]. A final reason to explicitly represent and infer causal structure in a perceptual model is that it may even be of intrinsic interest. For example, as we saw in the audio-visual context of Chapter 3, knowledge of structure corresponds to knowledge of "who said what" in a conversation [Hospedales and Vijayakumar, 2008].

4.3.4.2 Conclusions

In conclusion, the classical maximum likelihood integration theory, and other simpler theories for cue combination, appeared to fail with qualitative discrepancy in either the across or within-modal oddity detection experiments reported in [Hillis et al., 2002]. In this section, we have seen that these experiments can be explained quantitatively by a more accurate Bayesian observer model for this experiment, which exploits structure inference [Hospedales et al., 2007, Kording et al., 2007]. The structure inference approach therefore unifies the existing results for across and within-modality scenarios – and makes new testable predictions for further experiments. This result suggests that the brain may use a single principle for combining sensory information: including observations made both within and across modalities. Moreover, in addition to the audio-visual domain and direct estimation paradigm investigated by previously (Section 4.2, [Kording et al., 2007, Sato et al., 2007]), we have now provided evidence that structure inference occurs in combining visual-haptic as well as texture-disparity observations, and does so in a completely different oddity detection paradigm. The commonality of these results – across and within different types of modalities, and across different experimental paradigms – begins to suggest that structure inference may actually be a commonly evolved principle for combining perceptual information in the brain.

4.4 Bayesian Models of Human Perception: Discussion

4.4.1 Summary

In this chapter, we have applied our work on explicit probabilistic reasoning about multisensory data association to modelling recent experiments in human neurosciences.

Using this structure inference framework, we were able to provide a qualitative improvement in explanatory power over previous state of the art models for these experiments. Specifically, in the audio-visual domain, only by explicit representation of variable Bayesian network structure were we able to explain, for the first time, the phenomenon of unity perception and its interaction with location perception. In the visual-haptic and texture-disparity domains, this enabled us to explain the observed oddity detection rate contours – as well as perception of unity, although this was not measured in the experiment. It is significant that we were able to explain the results from two completely different sets of experiments – in different paradigms, and with different combinations of modalities – using a single probabilistic framework. This is important, because it suggests that human multisensory perception involves much more sophisticated probabilistic reasoning "under the hood" than had previously been thought based on classical sensor fusion models. Our results fit nicely into the emerging view of Bayesian inference as a general model for human perception, which includes likelihood and prior combination ([Kording and Wolpert, 2004a]), fusion of multiple independent cues ([Ernst and Banks, 2002, Alais and Burr, 2004]), and now cue combination under uncertain data association ([Hospedales and Vijayakumar, 2007, Kording et al., 2007]).

4.4.2 Task "Irrelevant" Modalities

Some authors have recently struggled to understand – in terms of the contemporary maximum likelihood integration framework – the "partial multisensory integration" observed when performing multisensory discrimination or detection tasks in the presence of a distracting task-irrelevant stimulus [Bresciani et al., 2006, Roach et al., 2006, Ernst, 2005]. In these cases, the task-irrelevant, distracting observation has an intermediate effect on the percept: It is not ignored entirely (even when subjects are explicitly instructed to ignore it), but it does not affect the final percept as strongly as would be predicted by maximum likelihood integration (eq. (1.1)). How can this be explained? In our structure inference framework (Figure 2.2), this is easy to explain. The brain would also consider the possibility that the "irrelevant" stimulus i is, actually, relevant (via the inference for association variable M_i (eqs. (2.2)-(2.4))). An *instruction* to the effect that it is irrelevant might reduce the strength of the prior probability π_{M_i} of association M_i in the context of this task, but may not reduce π_{M_i} to zero if the brain has learned over a lifetime of experience that these modalities are typically correlated

($\pi_{M_i} \gg 0$). This is because internal estimates of perceptual parameters may not be consciously accessible, and instead may be adapted over time with experience (e.g., [Ernst, 2007] illustrates learning potential correlation ($M_i > 0$) from previously uncorrelated ($M_i = 0$) cues). As such, for a previously correlated modality i, without extensive retraining that now $M_i = 0$, then in estimation of the stimulus, the brain weights the probability of M_is relevance (eqs. (2.2)-(2.4)). In the light of the prior and the evidence, this should produce an intermediate answer between integration and segregation (eq. (2.5)) as is observed ([Bresciani et al., 2006, Roach et al., 2006, Ernst, 2005]).

4.4.3 Additional Association Cues

In many contexts, there will be multiple types of cues to association. For example, in audio-visual perception (where we have focused on spatial alignment of the two observations) the *temporal synchrony* of the observations is also an important indicator of whether they should be unified or not – and hence, what location should be inferred. In the experiments discussed in this chapter, such additional cues have generally been controlled in order to focus on the effect of the main cues of interest.

Some recent work has examined the consequence of such additional cues on fusion. For example, fusion in visual-haptic size perception (as in [Ernst and Banks, 2002]) decreases with increasing *spatial* disparity [Gepshtein and Banks, 2003] – presumably due to decreasing posterior probability of unity. These additional dimensions of observations, and their effect on structure inference, can easily be included in the general framework proposed in this thesis. For example, similar and very recent work in this avenue has modelled the strong dependence of the ventriloquist effect (audio-visual localization) on temporal synchrony as well as spatial alignment of the audio-visual cues [Sato et al., 2007].

4.4.4 Relation to Other Neurosciences

What about the neuro-physiological implementation of these computations? There is some physiological evidence broadly in line with the behavioural psychophysics results and computational theories discussed in this chapter. In the audio-visual domain, for example, the superior colliculus (SC) is well studied, and known to respond to both audio and visual stimuli. Multi-sensory response strength is increased compared to unimodal response, when stimuli are spatially coincident, and decreased when stimuli are spatially discrepant [Witten and Knudsen, 2005]. Interestingly, recent re-

search suggests that AV multisensory interaction occurs very rapidly, and early in the processing hierarchy. The AV beep-flash counting illusion ([Shams et al., 2000], discussed in Section 2.2.3) has been the subject of much recent study. In experimental trails where the illusion occurs – and a beep creates the percept of a flash – increased response is visible in cortical area V1 under magnetic resonance imaging (MRI) [Watkins et al., 2006]. When the illusion occurs, visual event related potentials (VEPs) – qualitatively similar to those induced by true visual flashes – are also observed as rapidly as 170 ms after the auditory stimulus [Shams et al., 2001].

What kind of neural architecture might be able to perform the computations described in this chapter within the constraints of the known physiology (e.g., the latency of response discussed above)? Theoretical work on probabilistic population coding describes how neural populations could potentially encode probability distributions in their distributed firing statistics [Knill and Pouget, 2004, Pouget et al., 2003]. For many Bayesian computations, and in particular, those involved in multisensory fusion, we need to compute products of probability distributions. This has been shown theoretically for population codes, via common basis function layer [Deneve et al., 2001]. More recently, it has been shown that a population code exploiting the Poisson-like firing statistics of cortical neurons would be particularly well suited for performing such computations [Ma et al., 2006]. This would only involve a single linear operation, without the need for separate normalization steps, and hence potentially be very fast. It could also extend to encoding non-Gaussian distributions. Further experimental work is needed to confirm whether any of these proposed population coding models are actually implemented by biological neural networks.

Recent research into abstract human learning and reasoning has found Bayesian structure inference to be a surprisingly powerful explanatory model for higher level human cognition (e.g., categorical learning and reasoning [Tenenbaum et al., 2006]). The inferences in these tasks are performed over a much more general set of model structures than the very constrained sets we have considered in this chapter. Interestingly, however, despite the apparent dissimilarity between cognitive reasoning and subconscious perceptual processes, the mechanisms and principles are involved very similar. This suggests the intriguing possibility that probabilistic inference might be a broadly evolved mechanism for computation in the brain.

Chapter 5

Conclusions & Future Directions

5.1 Summary

In this thesis, we have considered the modelling of multisensory perception tasks relevant to both humans and machines. Prior work in multisensory machine perception has successfully applied classical sensor fusion theory to improve the performance of machine perception systems relative to those with a single sensory modality [Beal et al., 2003, Perez et al., 2004, Hershey and Movellan, 1999, Chen and Rui, 2004, Nefian et al., 2002, Serby et al., 2004]. However, as we saw in Section 3.3.1, this approach is limited because it is not robust to occlusion, sensor failure or other cause for outlying data. Although other heuristic schemes could be used to improve the robustness of these algorithms, without an explicit representation of data association, they are intrinsically limited in their ability to infer potentially important relational quantities such as "who said what?".

Classical sensor fusion theory has also been successfully applied by prior work in the human neurosciences to understand a broad range of phenomena in human multisensory perception. For a variety of senses and sensorimotor tasks, classical sensor fusion appears to provide a unifying explanation of the particular manner in which humans combine information from multiple senses [Ernst and Bulthoff, 2004, Ernst and Banks, 2002, van Beers et al., 2002, Alais and Burr, 2004, Jacobs, 1999, Hillis et al., 2004, Gepshtein and Banks, 2003]. Recent experiments however, have presented psychophysical tasks in which subjects discounted implausibly discrepant stimuli [Hairston et al., 2003, Wallace et al., 2004, Roach et al., 2006, Recanzone, 2003], and even made explicit use of the discrepancy to solve tasks [Hillis et al., 2002]. These results cannot be explained by classical sensor

fusion theory.

We modelled problems in multisensory perception with ambiguous data association using Bayesian networks with uncertain structure (Chapter 2). The generality of this formulation allowed us to apply it both to problems in machine learning of perceptual tasks and to problems in human psychophysics.

In machine perception (Chapter 3, [Hospedales and Vijayakumar, 2008, Hospedales et al., 2007]), we built a model to represent the high dimensional, intermittently correlated audio-visual data generated by a speaking and moving human. By performing inference in this model, we were able to perform audio-visual detection and tracking of human subjects. Moreover, we were able to do so in real time. By learning the parameters of the model with EM, we were able to learn the audio-visual mapping and the subject's appearance, which permitted unsupervised multisensory tracking. Ultimately, by approximating the full multi-target tracking problem with two independent models, the model was able to perform multi-target tracking and data association, answering the question of *who said what, where?* Unsupervised learning of such representations is an important step forward within the overarching theme of building more autonomous and capable cognitive robotics and AI systems.

In human perception (Chapter 4, [Hospedales and Vijayakumar, 2007]), we applied our framework for multisensory perception to analyse two recent experiments in the audio-visual [Hairston et al., 2003, Wallace et al., 2004] and visual-haptic [Hillis et al., 2002] domains. In the audio-visual domain (Section 4.2), we saw that by carefully considering the experimental task posed by [Hairston et al., 2003, Wallace et al., 2004], the appropriate model to use is a simplified version of - but in essence the same as - the machine perception model we developed in Chapter 3. By applying this model, we were able to explain numerous previously unexplained and counter-intuitive results in this series of experiments. In the visual-haptic domain (Section 4.3), we investigated the ground-breaking but perplexing experiments of [Hillis et al., 2002]. By applying our structure inference approach at two different levels (oddity and data association), we were able to provide the first complete explanation of the results of this experiment, which had not previously been fully qualitatively explained. These results – in such disparate modalities and experimental paradigms – are important because they suggest that apparently low level tasks in human mutisensory perception may involve much more sophisticated probabilistic reasoning under the hood than had previously been thought.

5.1.1 Summary of Contributions

The main contributions of our work are summarized as follows:

1. The interpretation of multisensory perception tasks with unknown data association as problems in Bayesian model inference (Chapter 2).

2. Specification of Bayesian models for various kinds of abstracted perceptual tasks (Chapter 2).

3. Application of our models to solve a large scale problem in machine learning for audio-visual scene understanding. This enabled real time detection, tracking, and speech segmentation; representing the quantities *who said what, where?* (Chapter 3)

   (a) Our model uses a *flexible appearance template*, rather than fixed domain specific knowledge such as facial features. As a result, it has broader applicability to other scenarios than most in the literature.

   (b) Our model performs *unsupervised learning* of the appearances and audio-visual mapping, and is therefore self calibrating. This is in contrast to typical scenarios requiring hand initialization of the target to be tracked, and extensive, time-consuming audio-visual calibration across the entire space.

4. Application of our models to understanding previously unexplained results in human audio-visual localisation and unity perception as consequences of Bayesian inference in the perceptual system (Chapter 4).

5. Application of our models to understanding previously unexplained results in human visual-haptic oddity detection as consequences of Bayesian inference in the perceptual system (Chapter 4).

This work in this thesis has led to the following publications:

- Structure Inference for Bayesian Multisensory Perception and Tracking
  Timothy Hospedales, Joel Cartwright and Sethu Vijayakumar
  Proc. International Joint Conference on Artificial Intelligence (IJCAI '07)
  Bibliographic Reference: [Hospedales et al., 2007]

- Structure Inference for Bayesian Multisensory Scene Understanding
  Timothy Hospedales and Sethu Vijayakumar
  IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear.
  Bibliographic Reference: [Hospedales and Vijayakumar, 2008]

- Bayesian Multisensory Oddity Detection
  Timothy Hospedales and Sethu Vijayakumar
  Neural Computation, submitted.
  Bibliographic Reference: [Hospedales and Vijayakumar, 2007]

5.2 Future Work

Machine Multisensory Perception In the context of our machine perception work, there are two avenues for future work that stand out. The first primarily deals with addressing weaknesses of the specific audio-visual formulation described in this thesis, and the second deals with unifying our system into a more complete sensorimotor framework.

Key limitations of the current formulation include some lack of robustness in the EM learning procedure (Section 3.3.2.3) and the lack of elegance of the solution in the current multi-target tracking framework (Section 3.3.3). Both limitations are largely due to the nature of the underlying TMG [Frey and Jojic, 2003] parametric framework chosen initially for this application. The lack of an explicit notion of visual layers and necessity of explaining the visual observation as a cyclically rotated template was the root cause of the EM learning limitations discussed in Section 3.3.2.3. For the same reason, it is difficult to represent the simultaneous effect of multiple sources on a given modality in the current formulation. To track multiple targets, we were therefore required to use entirely parallel models (Section 3.3.3). It would be more satisfying to define and perform inference on a single generative model for multiple targets (as was possible for the toy model in Section 2.2). Although this would incur more computational cost, the cost could potentially be dealt with, for example, via greedy [Williams and Titsias, 2004], variational [Jojic and Frey, 2001] or sampling [Williams et al., 2006] approximations.

A more interesting aim for future research is to unify our structure inference approach to multisensory perception within a complete sensorimotor loop. The topic of exploring with active perception has been of recent interest [Vijayakumar et al., 2001].

Some aspects of active perception require computations very similar to those investigated in this thesis. For example, in robot control it may be important to know the state (e.g., weight) of a manipulated object. In additional to passive sensor estimates (e.g., from vision), this can be estimated by observing the joint torques required to actively manipulate the object [Petkos and Vijayakumar, 2007]. Alternately, it can be estimated by observing the tactile forces applied during manipulation [Hoffmann et al., 2007]. Making the best estimate of the object state given all these observations is therefore an information fusion problem similar to those addressed in this thesis. Deciding *how* to actively manipulate the object in order to maximize the information gained in a multisensory context is an interesting open question.

Human Multisensory Perception In the context of studying human multisensory perception, there are also a number of interesting avenues for future work. Obviously there are further experiments, combinations of modalities, and tasks that can be investigated and compared to theory to discover the extent to which they are explainable by classical sensor fusion or the more sophisticated structure inference discussed in this thesis. A bigger interesting question is whether, in the cases where structure inference appears to apply, the brain is literally computing Bayesian model inference, approximating it (e.g., by ML, MAP or regularization), or simply applying a collection of clever task-specific heuristics which result in a similar response. Experimental designs should be conceived which can distinguish between these possibilities.

There is also the open question of parametrization. For Bayesian perception - even classical sensor fusion - the nervous system needs to have internal estimates of parameters such as the variance of its sensory apparatus. In our machine perception model, we used expectation-maximization to make ML estimates of these offline. How these parameters (some of which can even be updated rapidly and online [Jacobs and Fine, 1999]) are learnt by the nervous system is an interesting open question. Adaptation of cross-modal parameters, such as joint priors and association probabilities, in the light of multisensory observations is particularly topical [Ernst, 2007, van Beers et al., 2002].

Finally, there is the mechanistic question of the underlying neural implementation of Bayesian computations. As we discussed in Section 4.4.4, there has been recent theoretical work showing how various Bayesian computations could be performed by population code [Ma et al., 2006]. However, the actual existence of such population codes in the brain has yet to be verified.

Appendix A

Appendix

A.1 Audio-Visual Model Details

In this section we describe in more detail the inference and learning procedures used in the audio-visual model described in Chapter 3.

A.1.1 Factorial Hidden Markov Models

Standard hidden Markov models (HMMs, [Bishop, 2006b]) assume that the observations at every time are dependent on a single latent variable. In contrast, when there are multiple independent contributors to the observation at every time, FHMMs (Figure A.1, [Ghahramani and Jordan, 1997]) provide a useful representation.

Specifically, in a FHMM, the latent state $\mathbf{y}^t$ is composed of $i = 1..N$ factors y_i^t which have independent transition probabilities, so $p(\mathbf{y}^t|\mathbf{y}^{t-1}) = \prod_i^N p(y_i^t|y_i^{t-1})$. The observations $\mathbf{x}^t$ at each time-step potentially depends on all the latent factors $p(\mathbf{x}^t|\mathbf{y}^t)$. (When applied to data association specifically, the FHMM observation distribution $p(\mathbf{x}^t|\mathbf{y}^t)$ will use some of the latent factors y_i to gate the dependency of observations x_i on other factors y_j , (e.g., eq. (2.1)).

Inference and learning rules for FHMMs can be derived analogously to those for HMMs. The forward inference recursion $\alpha^t \triangleq p(\mathbf{x}^{1:t}, \mathbf{y}^t)$ – typically applied in tracking – can be derived as follows:

$$\begin{aligned} \alpha(\mathbf{x}^{1:t+1}, \mathbf{y}^{t+1}) &= \sum_{\mathbf{y}^t} p(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}, \mathbf{x}^{1:t}, \mathbf{y}^t), \\ &= \sum_{\mathbf{y}^t} p(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}|\mathbf{x}^t, \mathbf{y}^t) p(\mathbf{x}^{1:t}, \mathbf{y}^t), \end{aligned}$$

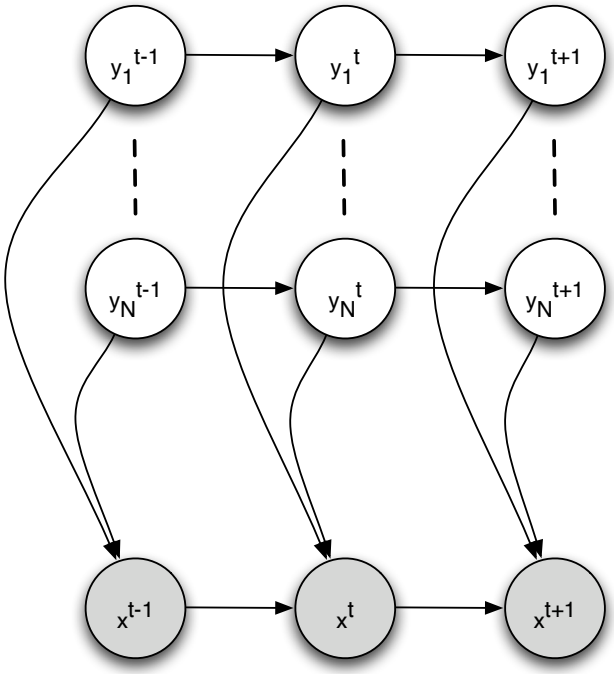


Figure A.1: A factorial hidden Markov model (FHMM).

$$\begin{aligned} &= \sum_{\mathbf{y}^t} p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1}) p(\mathbf{y}^{t+1}|\mathbf{y}^t) \alpha(\mathbf{x}^{1:t}, \mathbf{y}^t), \\ &= p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1}) \sum_{\mathbf{y}^t} \prod_{i=1}^{N} p(y_i^{t+1}|y_i^t) \alpha(\mathbf{x}^{1:t}, \mathbf{y}^t). \end{aligned} \tag{A.1}$$

The backward inference recursion $\gamma^t \triangleq p(\mathbf{y}^t|\mathbf{x}^{1:T})$ – which is required for learning – can be derived as follows:

$$\begin{aligned} \gamma(\mathbf{y}^t|\mathbf{x}^{1:T}) &= \sum_{\mathbf{y}^{t+1}} p(\mathbf{y}^t, \mathbf{y}^{t+1}|\mathbf{x}^{1:T}), \\ &= \sum_{\mathbf{y}^{t+1}} p(\mathbf{y}^t|\mathbf{y}^{t+1}, \mathbf{x}^{1:t}) p(\mathbf{y}^{t+1}|\mathbf{x}^{1:T}), \\ &= \sum_{\mathbf{y}^{t+1}} \frac{p(\mathbf{y}^t, \mathbf{y}^{t+1}, \mathbf{x}^{1:t})}{p(\mathbf{y}^{t+1}, \mathbf{x}^{1:t})} \gamma(\mathbf{y}^{t+1}|\mathbf{x}^{1:T}), \\ &= \sum_{\mathbf{y}^{t+1}} \frac{p(\mathbf{y}^{t+1}|\mathbf{y}^t) p(\mathbf{y}^t, \mathbf{x}^{1:t})}{\sum_{\mathbf{y}^t} p(\mathbf{y}^{t+1}|\mathbf{y}^t) p(\mathbf{y}^t, \mathbf{x}^{1:t})} \gamma(\mathbf{y}^{t+1}|\mathbf{x}^{1:T}), \\ &= \sum_{\mathbf{y}^{t+1}} \frac{\prod_{i=1}^{N} p(y_i^{t+1}|y_i^t) \alpha(\mathbf{x}^{1:t}, \mathbf{y}^t)}{\sum_{\mathbf{y}^t} \prod_{i=1}^{N} p(y_i^{t+1}|y_i^t) \alpha(\mathbf{x}^{1:t}, \mathbf{y}^t)} \gamma(\mathbf{y}^{t+1}|\mathbf{x}^{1:T}). \end{aligned} \tag{A.2}$$

This inference procedure is of $O(M^N)$ complexity for N Markov chains with M

states each, rendering it intractable for large numbers of chains. However, for tracking with data association in two modalities, (as discussed in Chapters 2 and 3), it is easily computable as only three chains are required.

A.1.2 Gaussian Linear Algebra Results

In this section, we show some results in linear algebra with Gaussians used throughout this thesis. Suppose we have two jointly Gaussian vectors $\mathbf{x},\mathbf{z}$ specified by the form $p(\mathbf{x},\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, where

$$\begin{aligned} p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x};\mathbf{G}\mathbf{z},\Psi), \\ p(\mathbf{z}) &= \mathcal{N}(\mathbf{z};\mu,\Phi). \end{aligned}$$

This form is common in perceptual inference problems, where observations $\mathbf{x}$ are described as being generated from a source $\mathbf{z}$ via some transformation $\mathbf{G}$ and sensor noise Ψ. The variability of source $\mathbf{z}$ is in turn described with mean μ and precision Φ. In this context we are interested in the data distribution $p(\mathbf{x})$ and the conditional distribution $p(\mathbf{z}|\mathbf{x})$ for inference. Since $\mathbf{x}$ and $\mathbf{z}$ are jointly Gaussian, these are both also Gaussian.

A.1.2.1 Conditional $p(\mathbf{z}|\mathbf{x})$

Since $p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z},\mathbf{x})$, we can determine the parameters of the conditional by rewriting the joint as a completed square in $\mathbf{z}$. Dropping the common $-\frac{1}{2}$ factor, the exponent of $p(\mathbf{x},\mathbf{z})$ is

$$\begin{aligned} & (\mathbf{x}-\mathbf{G}\mathbf{z})^T\Psi(\mathbf{x}-\mathbf{G}\mathbf{z}) + (\mathbf{z}-\mu)^T\Phi(\mathbf{z}-\mu), \\ = \;& \mathbf{x}^T\Psi\mathbf{x} - 2\mathbf{x}^T\Psi\mathbf{G}\mathbf{z} + \mathbf{z}^T\mathbf{G}^T\Psi\mathbf{G}\mathbf{z} + \mathbf{z}^T\Phi\mathbf{z} - 2\mu^T\Phi\mathbf{z} + \mu^T\Phi\mu, \\ = \;& \mathbf{z}^T(\mathbf{G}^T\Psi\mathbf{G}+\Phi)\mathbf{z} - 2\mathbf{z}^T(\mathbf{G}^T\Psi\mathbf{x}+\Phi\mu) + (\mathbf{x}^T\Psi\mathbf{x}+\mu^T\Phi\mu), \qquad \text{(A.3)} \\ = \;& (\mathbf{z}-\mu_{\mathbf{z}|\mathbf{x}})^T\Phi_{\mathbf{z}|\mathbf{x}}(\mathbf{z}-\mu_{\mathbf{z}|\mathbf{x}})^T - \mu_{\mathbf{z}|\mathbf{x}}\Phi_{\mathbf{z}|\mathbf{x}}\mu_{\mathbf{z}|\mathbf{x}} + \mathbf{x}^T\Psi\mathbf{x} + \mu^T\Phi\mu. \qquad \text{(A.4)} \end{aligned}$$

Where the mean $\mu_{\mathbf{z}|\mathbf{x}}$ and precision $\Phi_{\mathbf{z}|\mathbf{x}}$ of the conditional $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z};\mu_{\mathbf{z}|\mathbf{x}},\Phi_{\mathbf{z}|\mathbf{x}})$ can be determined by comparison with the exponent eq. (A.3) to be:

$$\begin{aligned} \Phi_{\mathbf{z}|\mathbf{x}} &= (\mathbf{G}^T\Psi\mathbf{G}+\Phi), & \text{(A.5)} \\ \mu_{\mathbf{z}|\mathbf{x}} &= \Phi^{-1}_{\mathbf{z}|\mathbf{x}}(\mathbf{G}^T\Psi\mathbf{x}+\Phi\mu). & \text{(A.6)} \end{aligned}$$

In our perceptual inference context, the posterior precision eq. (A.5) is now the sum of the likelihood (Ψ) and prior (Φ) precisions while the posterior mean eq. (A.6) is now a precision weighted sum of the likelihood ($\mathbf{x}$) and prior (μ) components.

A.1.2.2 Marginal $p(\mathbf{x})$

The marginal distribution $p(\mathbf{x})$ is determined from the joint $p(\mathbf{x},\mathbf{y})$ as $p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x},\mathbf{z}) = \int_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})p(\mathbf{x})$. From the exponent of the joint $p(\mathbf{x},\mathbf{z})$ in eq. (A.4), we see that integrating out the $\mathbf{z}$ factor leaves $-\mu_{\mathbf{z}|\mathbf{x}}\Phi_{\mathbf{z}|\mathbf{x}}\mu_{\mathbf{z}|\mathbf{x}} + \mathbf{x}^T\Psi\mathbf{x} + \mu^T\Phi\mu$, which we will rewrite as a completed square in $\mathbf{x}$ to find $p(\mathbf{x})$. Taking the $\mathbf{x}$ dependent terms, we have

$$\begin{aligned}
&-\mu_{\mathbf{z}|\mathbf{x}}\Phi_{\mathbf{z}|\mathbf{x}}\mu_{\mathbf{z}|\mathbf{x}} + \mathbf{x}^T\Psi\mathbf{x} \\
&= \mathbf{x}^T\Psi\mathbf{x} - \left((\mathbf{G}^T\Psi\mathbf{G}+\Phi)^{-1}(\mathbf{G}^T\Psi\mathbf{x}+\Phi\mu)\right)^T(\mathbf{G}^T\Psi\mathbf{G}+\Phi) \\
&\quad \cdot\left((\mathbf{G}^T\Psi\mathbf{G}+\Phi)^{-1}(\mathbf{G}^T\Psi\mathbf{x}+\Phi\mu)\right), \\
&= \mathbf{x}^T\left(\Psi - \Psi\mathbf{G}(\mathbf{G}^T\Psi\mathbf{G}+\Phi)^{-1}\mathbf{G}^T\Psi\right)\mathbf{x} - 2\mathbf{x}\Psi\mathbf{G}(\mathbf{G}^T\Psi\mathbf{G}+\Phi)^{-1}\Phi\mu, \\
&= \mathbf{x}^T(\Psi^{-1}+\mathbf{G}\Phi^{-1}\mathbf{G}^T)^{-1}\mathbf{x} - 2\mathbf{x}\Psi\mathbf{G}(\mathbf{G}^T\Psi\mathbf{G}+\Phi)^{-1}\Phi\mu, \qquad \text{(A.7)}
\end{aligned}$$

where in eq. (A.7), we have used the Woodbury identity[1] to simplify the precision term. The resulting Gaussian $p(\mathbf{x}) = \mathcal{N}(\mathbf{x};\mu_{\mathbf{x}},\Phi_{\mathbf{x}})$ can be identified as having mean and precision given by:

$$\mu_{\mathbf{x}} = \mathbf{G}\mu, \qquad \text{(A.8)}$$

$$\Phi_{\mathbf{x}} = (\Psi^{-1}+\mathbf{G}\Phi^{-1}\mathbf{G}^T)^{-1}. \qquad \text{(A.9)}$$

A.1.2.3 Multisensory Conditional $p(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2)$

The conditional distribution $p(\mathbf{z}|\mathbf{x})$ derived in Section A.1.2.1 generalizes straightforwardly to multiple observations. Consider making two observations $\mathbf{x}_1$ and $\mathbf{x}_2$ of $\mathbf{z}$ such that: $\mathbf{x}_1 \sim \mathcal{N}(\lambda_1\mathbf{G}_1\mathbf{z},\Psi_1)$, $\mathbf{x}_2 \sim \mathcal{N}(\lambda_2\mathbf{G}_2\mathbf{z},\Psi_2)$ and $\mathbf{z} \sim \mathcal{N}(\mu,\Phi)$. (Note that we have introduced an extra scaling factor λ_i of the observations i; it will be used in Chapter 3.) The conditional Gaussian is again determined by rewriting the joint product of Gaussians as a completed square in $\mathbf{z}$: $p(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2) \propto p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{z}) = p(\mathbf{x}_1|\mathbf{z})p(\mathbf{x}_2|\mathbf{z})p(\mathbf{z})$. Dropping the common $-\frac{1}{2}$ factor, the exponent of the joint is:

[1] $(A^{-1} - A^{-1}C(B^{-1}+C^TA^{-1}C)^{-1}C^TA^{-1}) \equiv (A+CBC^T)^{-1}$

$$
\begin{aligned}
&(\mathbf{x}_1-\lambda_1\mathbf{G}_1\mathbf{z})^T\Psi_1(\mathbf{x}_1-\lambda_1\mathbf{G}_1\mathbf{z})+(\mathbf{x}_2-\lambda_2\mathbf{G}_2\mathbf{z})^T\Psi_2(\mathbf{x}_2-\lambda_2\mathbf{G}_2\mathbf{z})+(\mathbf{z}-\mu)^T\Phi(\mathbf{z}-\mu)\\
&= \mathbf{z}^T(\lambda_1^2\mathbf{G}_1^T\Psi_1\mathbf{G}_1+\lambda_2^2\mathbf{G}_2^T\Psi_2\mathbf{G}_2+\Phi)\mathbf{z}-2\mathbf{z}^T(\lambda_1\mathbf{G}_1^T\Psi_1\mathbf{x}_1+\lambda_2\mathbf{G}_2^T\Psi_2\mathbf{x}_2+\Phi\mu)\\
&\quad+(\mathbf{x}_1^T\Psi_1\mathbf{x}_1+\mathbf{x}_2^T\Psi_2\mathbf{x}_2+\mu^T\Phi\mu), \qquad \text{(A.10)}\\
&= (\mathbf{z}-\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2})^T\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}(\mathbf{z}-\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2})^T-\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}^T\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}\\
&\quad+(\mathbf{x}_1^T\Psi_1\mathbf{x}_1+\mathbf{x}_2^T\Psi_2\mathbf{x}_2+\mu^T\Phi\mu), \qquad \text{(A.11)}
\end{aligned}
$$

where the mean $\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}$ and precision $\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}$ of the conditional $p(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2)=\mathcal{N}(\mathbf{z};\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2},\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2})$ can be determined by comparison with the exponent eq. (A.10) to be:

$$
\begin{aligned}
\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2} &= (\lambda_1^2\mathbf{G}_1^T\Psi_1\mathbf{G}_1+\lambda_2^2\mathbf{G}_2^T\Psi_2\mathbf{G}_2+\Phi), \qquad \text{(A.12)}\\
\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2} &= \Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}^{-1}(\lambda_1\mathbf{G}_1^T\Psi_1\mathbf{x}_1+\lambda_2\mathbf{G}_2^T\Psi_2\mathbf{x}_2+\Phi\mu). \qquad \text{(A.13)}
\end{aligned}
$$

In our perceptual inference context, the posterior precision eq. (A.12) is now the sum of the both of the likelihood (Ψ_1 and Ψ_2) and prior (Φ) precision terms, illustrating the increase in precision to be gained from making multiple independent observations. The posterior mean is now a precision weighted average of each of the observations $\mathbf{x}_1$ and $\mathbf{x}_2$, and the prior mean μ.

In the simplest scenario of scalar variables x_i and z without linear translations ($\lambda_i=1,\mathbf{G}_i=1$), then eqs. (A.12) and (A.13) are simplified to:

$$
\begin{aligned}
\phi_{z|x_1,x_2} &= (\Psi_1+\Psi_2+\Phi), \qquad \text{(A.14)}\\
\mu_{z|x_1,x_2} &= \phi_{z|x_1,x_2}^{-1}(\Psi_1x_1+\Psi_2x_2+\phi\mu), \qquad \text{(A.15)}
\end{aligned}
$$

where Ψ_1,Ψ_2,Φ and μ are now scalar.

A.1.2.4 Multisensory Marginal $p(\mathbf{x}_1,\mathbf{x}_2)$

The multisensory marginal distribution $p(\mathbf{x}_1,\mathbf{x}_2)$ is again determined from the joint $p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})$ as $p(\mathbf{x}_1,\mathbf{x}_2)=\int_{\mathbf{z}}p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{z})=\int_{\mathbf{z}}p(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2)p(\mathbf{x}_1,\mathbf{x}_2)$. From the exponent of the joint $p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{z})$ in eq. (A.11), we see that integrating out the $\mathbf{z}$ factor leaves $-\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}^T\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}+\mathbf{x}_1^T\Psi_1\mathbf{x}_1+\mathbf{x}_2^T\Psi_2\mathbf{x}_2+\mu^T\Phi\mu$, so that

$$
\begin{aligned}
-2\log p(\mathbf{x}_1,\mathbf{x}_2) &= -\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}^T\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}\mu_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}+\mathbf{x}_1^T\Psi_1\mathbf{x}_1+\mathbf{x}_2^T\Psi_2\mathbf{x}_2+\mu^T\Phi\mu\\
&\quad-\log\left(\frac{|\Psi_1|\,|\Psi_2|\,|\Phi|}{(2\pi)^2\left|\Phi_{\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2}\right|}\right). \qquad \text{(A.16)}
\end{aligned}
$$

In the simplest scenario of scalar variables x_i and z without linear translations ($\lambda_i = 1, \mathbf{G}_i = 1$), then eq. A.16 simplifies to

$$\begin{aligned} p(x_1,x_2) &= \sqrt{\frac{\Psi_1\Psi_2\Phi}{(2\pi)^2(\Psi_1+\Psi_2+\Phi)}} \\ &\cdot \exp\left(-\frac{1}{2}\frac{(x_1-x_2)^2\Psi_1\Psi_2+(x_1-\mu)^2\Psi_1\Phi+(x_2-\mu)^2\Psi_2\Phi}{\Psi_1+\Psi_2+\Phi}\right), \end{aligned} \tag{A.17}$$

where Ψ_1, Ψ_2, Φ and μ are now scalar. This is used in Chapter 4.

A.1.3 EM Parameter Updates

To optimize the model parameters $\theta = \{\lambda_{1,2}, \upsilon_{1,2}, \eta, \alpha, \beta, \omega, \pi_l, \mu, \Phi, \Psi, \Gamma, \Theta, \Omega, \pi_w, \pi_z, \gamma, \varepsilon, \sigma_{1,2}\}$, we make use of the posterior $p(H|D,\theta)$ over all hidden variables $H = \{\mathbf{a}, \mathbf{v}, \tau, l, \mathrm{W}, \mathrm{Z}\}_{t=1}^{T}$ (as computed in eq. (3.3)) and maximize the expected complete log likelihood of the data $\int_H p(H|D,\theta)\log p(H,D|\theta)$ with respect to the parameters θ:

$$\frac{\partial}{\partial\theta}\int_{\mathbf{a},\mathbf{v}}\sum_{\tau,l,W,Z} p(\{\mathbf{a},\mathbf{v},\tau,l,\mathrm{W},\mathrm{Z}\}_{t=1}^{T}|\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{y}\}_{t=1}^{T})\log p(\{\mathbf{a},\mathbf{v},\tau,l,\mathrm{W},\mathrm{Z},\mathbf{x},\mathbf{x}_2,\mathbf{y}\}_{t=1}^{T}).$$

The following sections list all the updates and some illustrative example derivations.

A.1.3.1 Video Appearance Model Updates

Consider optimizing for the video appearance parameter μ. The update is derived as follows:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\mu}\int_H p(H|D)\log p(H,D), \\ &= \sum_{t,Z,l}\int_{\mathbf{v}} p(\mathrm{Z},l,\mathbf{v}|D)\frac{\partial}{\partial\mu}\log\left(\mathcal{N}(\mathbf{v}|\mu,\Phi)^z\right), \\ &= \sum_{t,Z,l}\int_{\mathbf{v}} p(\mathbf{v}|l,Z,D)p(\mathrm{Z},l|D)z\Phi(\mathbf{v}-\mu), \\ \sum_{t,Z} p(\mathrm{Z}|D)z\mu &= \sum_{t,Z,l}\int_{\mathbf{v}} p(\mathbf{v}|l,Z,D)p(\mathrm{Z},l|D)z\mathbf{v}, \\ \sum_t p(z|D)\mu &= \sum_{t,l} p(l,z|D)\mu_{\mathbf{v}|\mathbf{y},l,z}, \\ \mu &= \frac{\sum_{t,l} p(l,z|D)\mu_{\mathbf{v}|\mathbf{y},l,z}}{\sum_t p(z|D)}. \end{aligned} \tag{A.18}$$

Here, we have identified $\int_{\mathbf{v}} p(\mathbf{v}|\mathbf{y},l,z)\mathbf{v}$ as the mean eq. (3.4) from the inference phase $\int_{\mathbf{v}} p(\mathbf{v}|\mathbf{y},l,z)\mathbf{v} = \mu_{\mathbf{v}|\mathbf{y},l,z}$. Explicit indexing by frame t was dropped for clarity.

The remaining video parameter updates are derived similarly and listed below:

$$\begin{aligned}
\mu &\leftarrow \frac{1}{N_z}\sum_{t,l} p(l^t,z^t|D^{1:T})\mu^t_{\mathbf{v}|\mathbf{y},l,z}, && \text{(A.19)}\\
\Phi^{-1} &\leftarrow \frac{1}{N_z}\sum_{t,l} p(l^t,z^t|D^{1:T}) \\
&\quad \cdot\mathrm{Diag}\left((\mu^t_{\mathbf{v}|\mathbf{y},l,z}-\mu)(\mu^t_{\mathbf{v}|\mathbf{y},l,z}-\mu)^T + (\nu^t_{\mathbf{v}|z})^{-1}\right), && \text{(A.20)}\\
\Psi^{-1} &\leftarrow \frac{1}{N_z N_{\mathbf{y}}}\sum_{t,l} p(l^t,z^t|D^{1:T}) \\
&\quad \cdot\mathrm{Tr}\left((\mathbf{y}^t-\mathbf{T}_l\mu^t_{\mathbf{v}|\mathbf{y},l,z})(\mathbf{y}^t-\mathbf{T}_l\mu^t_{\mathbf{v}|\mathbf{y},l,z})^T + (\nu^t_{\mathbf{v}|z})^{-1}\right), && \text{(A.21)}\\
\gamma &\leftarrow \frac{1}{N_z N_{\mathbf{y}}}\sum_{t} p(\overline{z}^t|D^{1:T})\sum_i \mathbf{y}^t_{[i]}, && \text{(A.22)}\\
\varepsilon^{-1} &\leftarrow \frac{1}{N_z N_{\mathbf{y}}}\sum_{t} p(\overline{z}^t|D^{1:T})(\mathbf{y}^t-\gamma)^2. && \text{(A.23)}
\end{aligned}$$

These updates make use of the sufficient statistics from the video inference, $\mu_{\mathbf{v}|\mathbf{y},l,z}$ and $\nu_{\mathbf{v}|z}$ (as computed in eqs. (3.4) and (3.5)). $N_z \triangleq \sum_t p(z^t|D^{1:T})$ and $N_{\overline{z}} \triangleq \sum_t p(\overline{z}^t|D^{1:T})$ are defined for convenience to be the total weight of associated and disassociated video frames respectively in the training sequence. $N_{\mathbf{y}}$ is the total number of pixels per frame and the inner product $\mathbf{x}^T\mathbf{x}$ is written as $\mathbf{x}^2$.

A.1.3.2 Audio Appearance Model Updates

Consider optimizing for the audio background noise precision parameter σ_1. The update is derived as follows:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\sigma_1}\int_H p(H|D)\log p(H,D), \\
&= \sum_{t,Z,l} p(\mathrm{W}|D)\frac{\partial}{\partial\sigma_1}\log\left(\mathcal{N}(\mathbf{x_1};0,\sigma)_1^{(1-w)}\right), \\
&= \sum_{t,Z,l}\int_{\mathbf{v}} p(\mathbf{v}|l,Z,D)p(\mathrm{Z},l|D)(1-w)(\mathbf{x}_1^T\mathbf{x}_1 - N_{\mathbf{x}}\sigma_1^{-1}), \\
\sum_{W,t} p(W|D)(1-w)N_{\mathbf{x}}\sigma_1^{-1} &= \sum_{W,t} p(W|D)(1-w)\mathbf{x}_1^T\mathbf{x}_1, \\
\sigma_1^{-1} &= \frac{\sum_t p(\overline{w}|D)\mathbf{x}_1^T\mathbf{x}_1}{N_{\mathbf{x}}\sum_t p(\overline{w}|D)}. && \text{(A.24)}
\end{aligned}$$

Here, explicit indexing by frame t was dropped for clarity and $N_{\mathbf{x}}$ is the total number of audio samples per frame. The remaining audio parameter updates are derived similarly and listed below:

$$\lambda_1 \leftarrow \frac{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})\mathbf{x}_1^T \mu_{\mathbf{a}|\mathbf{x}_1,\tau,w}}{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mu^t_{\mathbf{a}|\mathbf{x}_1,\tau,w})^2 + N_w \mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w})}, \tag{A.25}$$

$$\lambda_2 \leftarrow \frac{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})\mathbf{x}_2^T \mathbf{T}_\tau \mu_{\mathbf{a}|\mathbf{x}_2,\tau,w}}{\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mu^t_{\mathbf{a}|\mathbf{x}_2,\tau,w})^2 + N_w \mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w})}, \tag{A.26}$$

$$\upsilon_1^{-1} \leftarrow \frac{1}{N_w N_{\mathbf{x}}} \left(\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mathbf{x}_1^t - \lambda_1 \mu^t_{\mathbf{a}|\mathbf{x}_1,\tau,w})^2 + N_w \lambda_1^2 \mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w}) \right), \tag{A.27}$$

$$\upsilon_2^{-1} \leftarrow \frac{1}{N_w N_{\mathbf{x}}} \left(\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mathbf{x}_2^t - \lambda_2 \mathbf{T}_\tau \mu^t_{\mathbf{a}|\mathbf{x}_2,\tau,w})^2 + N_w \lambda_2^2 \mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w}) \right), \tag{A.28}$$

$$\eta^{-1} \leftarrow \frac{1}{N_w N_{\mathbf{x}}} \left(\sum_{t,\tau} p(\tau^t, w^t|D^{1:T})(\mu^t_{\mathbf{a}|\mathbf{x},\tau,w})^2 + N_w \mathrm{Tr}(\nu^{-1}_{\mathbf{a}|w}) \right), \tag{A.29}$$

$$\sigma_1^{-1} \leftarrow \frac{1}{N_{\overline{w}} N_{\mathbf{x}}} \sum_t p(\overline{w}^t|D^{1:T})(\mathbf{x}_1^t)^2, \tag{A.30}$$

$$\sigma_2^{-1} \leftarrow \frac{1}{N_{\overline{w}} N_{\mathbf{x}}} \sum_t p(\overline{w}^t|D^{1:T})(\mathbf{x}_2^t)^2. \tag{A.31}$$

Here we make use of the sufficient statistics from the audio inference, $\mu_{\mathbf{a}|\mathbf{x},\tau,w}$ and $\nu_{\mathbf{a}|w}$ as computed in eqs. (3.6)-(3.7). The full posterior over delay, location and association $p(\tau^t, l^t, \mathrm{W}^t, Z^t|D^{1:T}) = p(\tau^t|l^t, \mathrm{W}^t, \mathrm{D}^{1:T}) p(l^t, \mathrm{W}^t, Z^t|D^{1:T})$ is also used (as computed by eqs. (3.9) and (3.20)). $N_w = \sum_t p(w^t|D^{1:T})$ and $N_{\overline{w}} = \sum_t p(\overline{w}^t|D^{1:T})$ are defined for convenience to be the total weight of associated and disassociated audio frames respectively in the training sequence.

A.1.3.3 Multimodal Updates

Suppressing indexing by time t for clarity, the parameters of the audio-visual link are updated as follows:

$$\beta \leftarrow \frac{1}{N_w} \left(\sum_{\tau,l,t} \tau \cdot Q^t_{t,l,w} - \alpha \sum_{\tau,l,t} l \cdot Q^t_{t,l,w} \right), \tag{A.32}$$

$$\alpha \leftarrow \frac{\sum_{\tau,t,l} Q^t_{t,l,w} \left(l\tau - l \frac{1}{N_w} \sum_{\tau,t,l} Q^t_{t,l,w} \tau \right)}{\sum_{\tau,t,l} Q^t_{t,l,w} \cdot l^2 - \left(\sum_{\tau,t,l} Q^t_{t,l,w} \cdot l \cdot \left(\frac{1}{N_w} \sum_{\tau,t,l} Q^t_{t,l,w} l \right) \right)}, \tag{A.33}$$

$$\omega^{-1} \leftarrow \frac{1}{N_w} \sum_{\tau,t,l} Q^t_{t,l,w} \left(\tau^2 - 2\tau\alpha l - 2\tau\beta + \alpha^2 l^2 + 2\alpha l\beta + \beta^2 \right), \tag{A.34}$$

where for brevity we define the notation $Q^t_{t,l,w} \triangleq p(\tau^t, l^t, w^t | D^{1:T})$ for the posterior over time-delay, location and audibility.

A.1.3.4 Markov chain updates

To compute the updates for the Markov chain parameters, we make use of the sufficient statics α eq (2.6) and γ eq. (2.7) from the inference as follows:

$$\Gamma_{[i,j]} \leftarrow \frac{\sum_t p(l^t, l^{t+1} | D^{1:T})_{[i,j]}}{\sum_t \gamma(l^t)_{[i]}}, \tag{A.35}$$

$$\Theta_{[i,j]} \leftarrow \frac{\sum_t p(W^t, W^{t+1} | D^{1:T})_{[i,j]}}{\sum_t \gamma(W^t)_{[i]}}, \tag{A.36}$$

$$\Omega_{[i,j]} \leftarrow \frac{\sum_t p(Z^t, Z^{t+1} | D^{1:T})_{[i,j]}}{\sum_t \gamma(Z^t)_{[i]}}, \tag{A.37}$$

where $p(l^t, l^{t+1} | D^{1:T}) = \frac{\alpha(l^t) p(D^{t+1} | l^{t+1}) \gamma(l^t) \Gamma_{[l^t, l^{t+1}]}}{\alpha(l^{t+1})}$.

A.1.4 FFT Speedup Equations

A.1.4.1 Inference

The following FFT based computations (derived along the lines described in Section 3.2.4.1) were used to speed up inference (eqs. (3.12), (3.16) and (3.13) respectively):

$$\log p(\tau | w, l, \mathbf{x}_1, \mathbf{x}_2) = \log p(\tau | l) + \lambda_1 \lambda_2 \upsilon_1 \upsilon_2 \nu^{-1}_{\mathbf{a}|w} \mathrm{Corr}[\mathbf{x}_1, \mathbf{x}_2] + \log K, \tag{A.38}$$

$$\begin{aligned} 2\log p(\mathbf{x}_1, \mathbf{x}_2 | \tau, w) = & \\ & v^{-1}_{\mathbf{a}|w} \sum_i \left(\lambda_1^2 \upsilon_1^2 \mathbf{x}_1[i]^2 + 2\lambda_1 \lambda_2 \upsilon_1 \upsilon_2 \mathrm{Corr}[\mathbf{x}_1, \mathbf{x}_2] + \lambda_2^2 \upsilon_2^2 \mathbf{x}_2[i]^2 \right) \\ & -(\mathbf{x}_2^T \upsilon_2 \mathbf{x_2} + \mathbf{x}_1^T \upsilon_1 \mathbf{x}_1) + \log \left(\frac{|\upsilon_1| |\upsilon_2| |\eta|}{4\pi^2 \left| \nu_{\mathbf{a}|w} \right|} \right) \end{aligned} \tag{A.39}$$

$$\begin{aligned} -2\log(\mathbf{y} | l, z) = & \log \left(\left| 2\pi \nu_{\mathbf{y}|l,z} \right| \right) + \mathrm{Corr}\left[\mathbf{y}^2, (\phi^{-1} + \psi^{-1})^{-1} \right] \\ & -2\mathrm{Corr}\left[\mathbf{y}, \mu * (\phi^{-1} + \psi^{-1})^{-1} \right] + \mu^T \mathbf{T}_l^T \nu_{\mathbf{y}|l,z} \mathbf{T}_l \mu. \end{aligned} \tag{A.40}$$

Here we make use of the audio posterior precision (3.7) $\nu_{\mathbf{a}|w} = \eta + \lambda_1^2 \upsilon_1 + \lambda_2^2 \upsilon_2$ and the video data likelihood precision matrix (3.14), $\nu_{\mathbf{y}|l,z} = (\Psi^{-1} + \mathbf{T}_l \Phi^{-1} \mathbf{T}_l^T)^{-1}$. ϕ and ψ represent the diagonal elements of Φ and Ψ respectively. $*$ denotes element-wise multiplication.

A.1.4.2 Learning

The following FFT based computations (derived along the lines described in Section 3.2.4.1) were used to improve the efficiency of EM updates (eqs. (A.19), (A.20) and (A.21) respectively):

$$\mu \leftarrow \frac{1}{N_z}(\Phi+\Psi)^{-1}\sum_t^T\left(q_{lz}^t\Phi\mu+\psi\text{Conv}\left[q_{lz}^t,\mathbf{y}^t\right]\right), \tag{A.41}$$

$$\phi^{-1} \leftarrow$$

$$v_{\mathbf{v}|z}^{-1}\frac{1}{N_z}\sum_t^T\left(q_z^t+v_{\mathbf{v}|z}^{-1}\Psi^2\left(\text{Conv}\left[q_{lz}^t,\mathbf{y}^t*\mathbf{y}^t\right]-\mu*\text{Conv}\left[q_{lz}^t,\mathbf{y}^t\right]+q_z^t\mu*\mu\right)\right), \tag{A.42}$$

$$N_{\mathbf{y}}N_z\Psi^{-1} \leftarrow \sum_t q_z^t\text{Tr}\left[v_{\mathbf{v}|z}^{-1}\right]+\sum_t\text{Tr}\left[v_{\mathbf{v}|z}^{-2}\mathbf{y}^t\left(\text{Corr}\left[\phi*\phi,q_{lz}^t\right]^T\right.\right.$$
$$\left.\left.-2\mathbf{y}^t\text{Corr}\left[\phi*\mu,q_{lz}^t\right]^T+\mathbf{y}^T\text{Corr}\left[\phi*\phi*\mu*\mu,q_{lz}^t\right]^T\right)\right]. \tag{A.43}$$

Here, we define the notation $q_{lz}^t \triangleq p(l^t,z^t|D^{1:T})$ and $q_z^t \triangleq p(z^t|D^{1:T})$ for the location and visibility posteriors.

A.2 Oddity Detection Details

A.2.1 Complete Results

In this section, we summarize the predictions of our model and the experimental data for eight of the reported scenarios in [Hillis et al., 2002]. In Figure A.2(a)-(d), the across-modality predictions are shown. The unimodal variances (determined by the red lines) and prior probability of fusion (fit to the data) vary across subjects and enable accurate prediction of the multimodal detection contours in almost every case.

Our approach required one compromise in modelling power compared to [Hillis et al., 2002]. One of the particular within-modality cues used in this experiment, texture based slant perception, exhibits the unusual property of being perceived with less variance as a function of the actual stimulus slant [Knill, 1998]. So the noisy observations should ideally be modelled as $x_t \sim \mathcal{N}(y_t,\sigma_t(y_t))$ where $\sigma_t(y_t)$ is a decreasing function of the absolute magnitude of y_t. This is why the human data in the within-modal experiments (Figure A.2(e)-(h)) takes a more curved shape than the across-modal experiments (Figure A.2(a)-(d)). This is particularly apparent when the standard stimulus is taken to not be at $(0,0)$ because then the variances of observations above and below the standard are not even symmetrical (Figure A.2(f)-(h)).

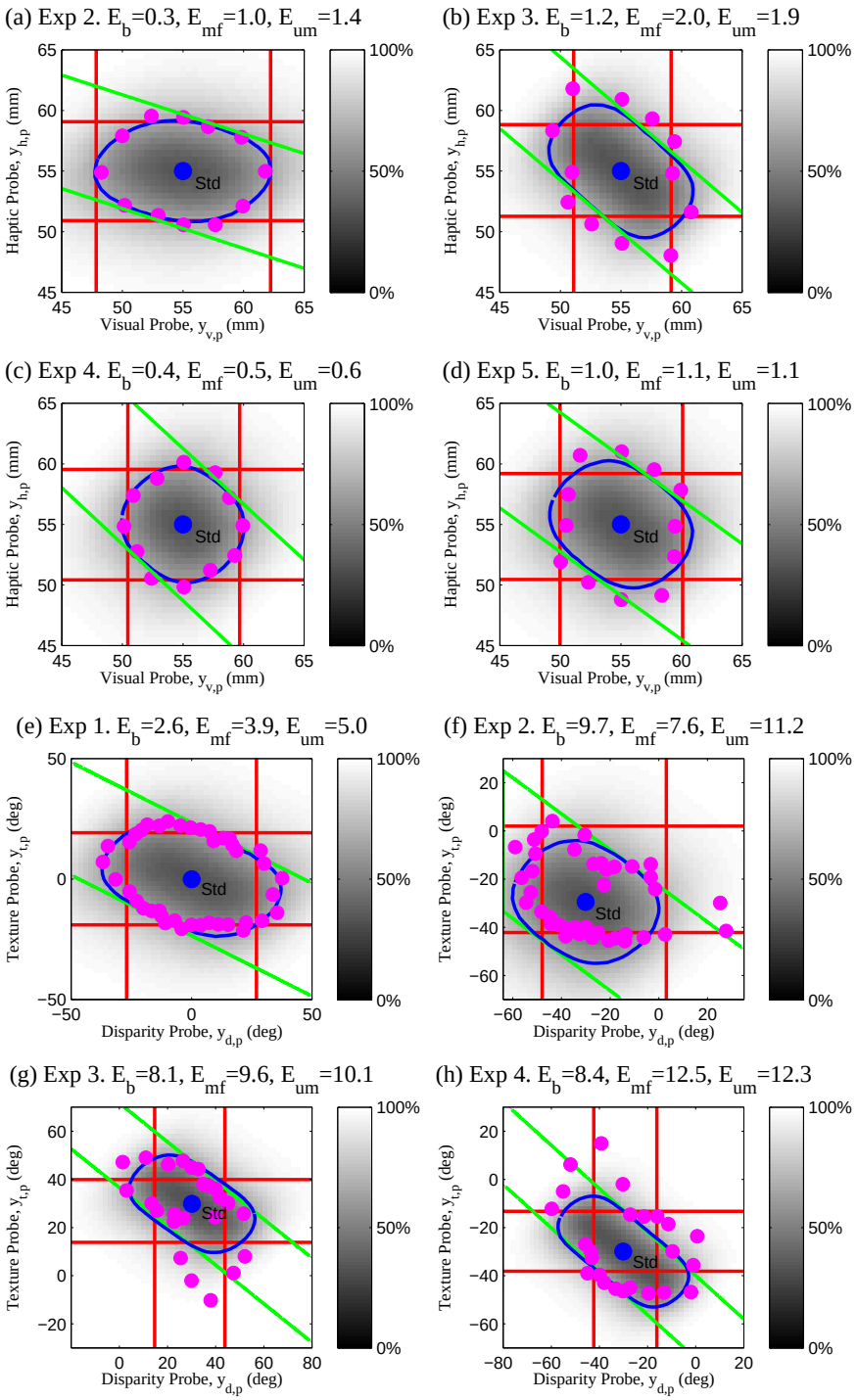


Figure A.2: Oddity detection rate predictions for ideal Bayesian observer using variable structure model (grey-scale background). Oddity detection rate threshold contours for the Bayesian model (blue lines), mandatory fusion model (green lines) and unimodal model (red lines) are shown along with human thresholds (magenta points). (a-d) Visual-haptic condition. (e-h) Texture-disparity condition. Chance=33%. Contour root mean squared error is given for; E_b: Bayesian model, E_{mf}: sequential fused estimate and unimodal model, E_{um}: sequential unimodal model.

The simpler overall model in [Hillis et al., 2002] allowed them to model the dependence of $\sigma_t(y_t)$ on y_t numerically, while still retaining computational tractability (hence, the curved green prediction lines in Figure 4.9(b)). Note, however, that this introduces additional free parameters in the function $\sigma_t(y_t)$, which they tuned to fit the data.

In our approach, we were not able to incorporate variable and asymmetric variance while retaining analytical and computational tractability of the model, and we simply assumed it was constant and symmetric. Hence, the fits of our model to the within-modal data with asymmetric variance (Figure A.2(f)-(h)) do not have the same quantitative accuracy as for the other experiments (Figure A.2(a)-(e)). Nevertheless, even in these cases, the essential slightly elongated region of non-detection along the cues-discordant axis is still captured albeit without the curve related to $\sigma_t(y_t)$. In future work, this limitation could be potentially addressed while retaining the same general framework by including a parametrised dependency $\sigma_t(y_t)$ in the generative model and applying a sampling, rather than analytical, approach to integrating the latent variables $\{y_p, y_{t,p}, y_{d,p}, y_s\}$ in eq. (4.11).

A.2.2 Oddity Inference

The model likelihoods can be determined by simple integrals of Gaussian products. We assume all the observations are distributed normally given the source $x_{h,i} \sim \mathcal{N}(y, \sigma_h^2)$ and $x_{v,i} \sim \mathcal{N}(y, \sigma_v^2)$, and that the subject's prior belief about the source locations is represented by $y_s \sim \mathcal{N}(\mu_s, \sigma_y)$ and $y_p \sim \mathcal{N}(\mu_s, \sigma_y)$. Then, the model likelihood $p(\{x_{h,i}, x_{v,i}\}_{i=1}^3 | p, \theta)$ can be determined by integrals of Gaussian products. (See Appendix A.1.2.4 for details.) Eqs. (A.44)-(A.48) detail specific solutions as follows:

$$\begin{aligned}
p(\{x_{h,i}, x_{v,i}\}_{i=1}^3 | p) &= \int_{y_s, y_p} p(\{x_{h,i}, x_{v,i}\}_{i\in\{1,2,3\}\setminus p} | y_s, p) p(x_{h,p}, x_{v,p} | y_p, p) p(y_p) p(y_s), \\
&= p_s(\{x_{h,i}, x_{v,i}\}_{i\in\{1,2,3\}\setminus p} | p) p_p(x_{h,p}, x_{v,p} | p), \qquad \text{(A.44)} \\
p_s(\{x_{h,i}, x_{v,i}\}_{i\in\{1,2,3\}\setminus p} | p) &= \int_{y_s} \prod_{i\in\{1,2,3\}\setminus p} \prod_{j=h,v} \mathcal{N}(x_{j,i} | y_s) \mathcal{N}(y_s), \qquad \text{(A.45)} \\
p_s(\{x_{h,i}, x_{v,i}\}_{i\in\{1,2\}} | p = 3) &\propto \\
\exp -\frac{1}{2} &\left[-\frac{(x_{h,1} + x_{h,2})\rho_h + (x_{v,1} + x_{v,2})\rho_v}{2\rho_h + 2\rho_v + \rho_y} + (x_{h,1}^2 + x_{h,2}^2)\rho_h + (x_{v,1}^2 + x_{v,2}^2)\rho_v \right], \qquad \text{(A.46)} \\
p_p(x_{h,p}, x_{v,p} | p) &= \int_{y_p} \mathcal{N}(x_{h,p} | y_s) \mathcal{N}(x_{v,p} | y_s) \mathcal{N}(y_p), \qquad \text{(A.47)} \\
p_p(x_{h,p}, x_{v,p} | p = 3) &\propto
\end{aligned}$$

$$\exp -\frac{1}{2}\frac{1}{(\rho_h+\rho_v+\rho_y)}\left(-2x_{h,3}x_{v,3}\rho_h\rho_v+x_{v,3}^2\rho_v(\rho_h+\rho_y)+x_{h,3}^2\rho_h(\rho_v+\rho_y)\right). \quad \text{(A.48)}$$

Note that for simplicity, we use precisions $\rho_i=\sigma_i^{-2}$ rather than variances σ_i^2, assume that object three is odd $(p=3)$ and that $\mu_s=0$. All distributions are assumed to be conditional on the parameters θ.

A.2.3 Oddity Inference with Variable Structure

Conditioned on the causal structure $C\in\{c,\bar{c}\}$ as well as the model (oddity) p, all the likelihoods factor and are still determined by integrals of Gaussian products. To compute the model posterior, we integrate out the binary causal structure variable C numerically. With the same assumptions as in Section A.2.2, the likelihoods are as follows:

$$\begin{aligned}
p(\{x_{h,i},x_{v,i}\}_{i=1}^3|p) &= \\
&\sum_C\int p(\{x_{h,i},x_{v,i}\}_{i=1}^3,y_s,y_p,y_{h,p},y_{v,p},C,|p)dy_s dy_p dy_{h,p}dy_{v,p},\\
&= p_s(\{x_{h,i},x_{v,i}\}_{i\in\{1,2,3\}\setminus p}|p,\theta)p_p(x_{h,p},x_{v,p}|p), \quad \text{(A.49)}\\
p_s(\{x_{h,i},x_{v,i}\}_{i\in\{1,2,3\}\setminus p}|p) &= \int_{y_s}\prod_{i\in\{1,2,3\}\setminus p}\prod_{j=h,v}\mathcal{N}(x_{j,i}|y_s)\mathcal{N}(y_s),\\
p_s(\{x_{h,i},x_{v,i}\}_{i\in\{1,2,3\}\setminus p}|p=3) &\propto \\
\exp -\frac{1}{2}&\left[-\frac{(x_{h,1}+x_{h,2})\rho_h+(x_{v,1}+x_{v,2})\rho_v}{2\rho_h+2\rho_v+\rho_y}+(x_{h,1}^2+x_{h,2}^2)\rho_h+(x_{v,1}^2+x_{v,2}^2)\rho_v\right], \quad \text{(A.50)}\\
p_p(x_{h,p},x_{v,p}|p) &= \int_{y_p}p(x_{h,p},x_{v,p},y_p|p,c)p(c)\\
&\quad+\int_{y_{h,p}y_{v,p}}p(x_{h,p},x_{v,p},y_{h,p},y_{v,p}|p,\bar{c})p(\bar{c}), \quad \text{(A.51)}\\
p_p(x_{h,p},x_{v,p}|p,c) &= \int_{y_p}\mathcal{N}(x_{h,p}|y_s,c)\mathcal{N}(x_{v,p}|y_s,c)\mathcal{N}(y_p|c),\\
p_p(x_{h,p},x_{v,p}|p=3,c) &\propto \\
\exp -\frac{1}{2}&\frac{1}{(\rho_h+\rho_v+\rho_y)}\left(-2x_{h,3}x_{v,3}\rho_h\rho_v+x_{v,3}^2\rho_v(\rho_h+\rho_y)+x_{h,3}^2\rho_h(\rho_v+\rho_y)\right), \quad \text{(A.52)}\\
p_p(x_{h,p},x_{v,p}|p,\bar{c}) &= \int_{y_{h,p}y_{v,p}}\mathcal{N}(x_{h,p}|y_{h,p},\bar{c})\mathcal{N}(x_{v,p}|y_{v,p},\bar{c})\mathcal{N}(y_{h,p}|\bar{c})\mathcal{N}(y_{v,p}|\bar{c}),\\
p(x_{h,p},x_{v,p}|p=3,\bar{c}) &= \mathcal{N}(x_{h,3};0,(\rho_h^{-1}+\rho_v^{-1})^{-1})\mathcal{N}(x_{v,3};0,(\rho_v^{-1}+\rho_v^{-1})^{-1}) \quad \text{(A.53)}
\end{aligned}$$

Bibliography

[cle, 2006] (2006). CLEAR 2006 evaluation and workshop campaign. http://www.clear-evaluation.org/.

[Al-Hames et al., 2006] Al-Hames, M., Hörnler, B., Scheuermann, C., and Rigoll, G. (2006). Using audio, visual, and lexical features in a multi-modal virtual meeting director. In *MLMI 2006, 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.

[Alais and Burr, 2004] Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol*, 14(3):257–262.

[Bar-Shalom et al., 2005] Bar-Shalom, Y., Kirubarajan, T., and Lin, X. (2005). Probabilistic data association techniques for target tracking with applications to sonar, radar and eo sensors. *IEEE Aerospace and Electronic Systems Magazine*, 20(8):37–56.

[Bar-Shalom and Tse, 1975] Bar-Shalom, Y. and Tse, E. (1975). Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11:451–460.

[Battaglia et al., 2003] Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1391–1397.

[Beal et al., 2003] Beal, M. J., Jojic, N., and Attias, H. (2003). A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836.

[Beierholm et al., 2007] Beierholm, U., Kording, K., Shams, L., and Ma, W. J. (2007). Comparing bayesian models for multisensory cue combination without mandatory integration. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*. MIT Press.

[Bilmes, 2000] Bilmes, J. (2000). Dynamic bayesian multinets. In *UAI*.

[Bishop, 2006a] Bishop, C. M. (2006a). *Pattern Recognition and Machine Learning*. Springer.

[Bishop, 2006b] Bishop, C. M. (2006b). *Pattern Recognition and Machine Learning*, chapter Sequential Data, pages 605–652. Springer.

[Blake and Isard, 1997] Blake, A. and Isard, M. (1997). The condensation algorithm - conditional density propagation and applications to visual tracking. *Advances in Neural Information Processing Systems*, 9:361–368.

[Boutilier et al., 1996] Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context-specific independence in bayesian networks. In *Uncertainty in Artificial Intelligence 1996*.

[Bresciani et al., 2006] Bresciani, J.-P., Dammeier, F., and Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *J Vis*, 6(5):554–564.

[Checka et al., 2004] Checka, N., Wilson, K., Siracusa, M., and Darrell, T. (2004). Multiple person and speaker activity tracking with a particle filter. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 5, pages V–881–4vol.5.

[Chen and Rui, 2004] Chen, Y. and Rui, Y. (2004). Real-time speaker tracking using particle filter sensor fusion. *Proceedings of the IEEE*, 92(3):485–494.

[Clark and Yuille, 1990] Clark, J. J. and Yuille, A. L. (1990). *Data Fusion for Sensory Information Processing Systems*. Springer.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

[Deneve et al., 2001] Deneve, S., Latham, P. E., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nat Neurosci*, 4(8):826–831.

[Ernst, 2005] Ernst, M. (2005). *Perception of the human body from the inside out*, chapter A Bayesian view on multimodal cue integration, pages 105–131. Oxford University Press.

[Ernst, 2007] Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, 7:1–14.

[Ernst and Banks, 2002] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415:429–433.

[Ernst and Bulthoff, 2004] Ernst, M. O. and Bulthoff, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn Sci*, 8(4):162–169.

[Fisher and Darrell, 2004] Fisher, J.W., I. and Darrell, T. (2004). Speaker association with signal-level audiovisual fusion. *Multimedia, IEEE Transactions on*, 6(3):406–413.

[Fortmann et al., 1983] Fortmann, T. E., Bar-Shalom, Y., and Scheffe, M. (1983). Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Engineering*, 8:173–184.

[Frey and Jojic, 2003] Frey, B. and Jojic, N. (2003). Transformation-invariant clustering using the em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):1–17.

[Frey and Jojic, 2005] Frey, B. J. and Jojic, N. (2005). A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1392 – 1416.

[Gatica-Perez et al., 2007] Gatica-Perez, D., Lathoud, G., Odobez, J.-M., and McCowan, I. A. (2007). Audio-visual probabilistic tracking of multiple speakers in meetings. *IEEE Trans. on Audio, Speech, and Language Processing*, 15:601–616.

[Geiger and Heckerman, 1996] Geiger, D. and Heckerman, D. (1996). Knowledge representation and inference in similarity networks and bayesian multinets. *Artificial Intelligence*, 82:45–74.

[Gepshtein and Banks, 2003] Gepshtein, S. and Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Curr Biol*, 13(6):483–488.

[Ghahramani and Jordan, 1997] Ghahramani, Z. and Jordan, M. (1997). Factorial hidden markov models. *Machine Learning*, 29:245–273.

[Hain et al., 2005] Hain, T., Dines, J., Garau, G., Karafiat, M., Moore, D., Wan, V., Ordelman, R., and Renals, S. (2005). Transcription of conference room meetings: an investigation. In *Transcription of Conference Room Meetings: an Investigation*.

[Hairston et al., 2003] Hairston, W. D., Wallace, M. T., Vaughan, J. W., Stein, B. E., Norris, J. L., and Schirillo, J. A. (2003). Visual localization ability influences cross-modal bias. *J Cogn Neurosci*, 15(1):20–29.

[Hershey and Movellan, 1999] Hershey, J. and Movellan, J. R. (1999). Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems*.

[Hillis et al., 2002] Hillis, J. M., Ernst, M. O., Banks, M. S., and Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science*, 298(5598):1627–1630.

[Hillis et al., 2004] Hillis, J. M., Watt, S. J., Landy, M. S., and Banks, M. S. (2004). Slant from texture and disparity cues: optimal cue combination. *J Vis*, 4(12):967–992.

[Hoffmann et al., 2007] Hoffmann, H., Petkos, G., Bitzer, S., and Vijayakumar, S. (2007). Sensor assisted adaptive motor control under continuously varying context. In *Proc. International Conference on Informatics in Control, Automation and Robotics (ICINCO '07)*.

[Hospedales et al., 2007] Hospedales, T., Cartwright, J., and Vijayakumar, S. (2007). Structure inference for bayesian multisensory perception and tracking. In *International Joint Conference on Artificial Intelligence 2007*.

[Hospedales and Vijayakumar, 2007] Hospedales, T. and Vijayakumar, S. (2007). Bayesian multisensory oddity detection. *Neural Computation, submitted*.

[Hospedales and Vijayakumar, 2008] Hospedales, T. and Vijayakumar, S. (2008). Structure inference for bayesian multisensory scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear*.

[Jacobs, 1999] Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Res*, 39(21):3621–3629.

[Jacobs and Fine, 1999] Jacobs, R. A. and Fine, I. (1999). Experience-dependent integration of texture and motion cues to depth. *Vision Res*, 39(24):4062–4075.

[Jojic and Frey, 2001] Jojic, N. and Frey, B. (2001). Learning flexible sprites in video layers. In *Computer Vision and Pattern Recognition, 2001*, volume 1.

[Jojic et al., 2000] Jojic, N., Petrovic, N., Frey, B., and Huang, T. (2000). Transformed hidden markov models: estimating mixture models of images and inferring spatial transformations in video sequences. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 26–33vol.2.

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82:35–45.

[Kersten et al., 2004] Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, 55:271–304.

[Knill, 1998] Knill, D. C. (1998). Discrimination of planar surface slant from texture: human and ideal observers compared. *Vision Res*, 38(11):1683–1711.

[Knill and Pouget, 2004] Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*, 27(12):712–719.

[Knill and Richards, 1996] Knill, D. C. and Richards, W., editors (1996). *Perception as Bayesian Inference*. Cambridge University Press.

[Kording and Tenenbaum, 2006] Kording, K. and Tenenbaum, J. B. (2006). Causal inference in sensorimotor integration. In *Advances in Neural Information Processing Systems 19*.

[Kording et al., 2007] Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, 2(9):e943.

[Kording and Wolpert, 2004a] Kording, K. P. and Wolpert, D. M. (2004a). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247.

[Kording and Wolpert, 2004b] Kording, K. P. and Wolpert, D. M. (2004b). The loss function of sensorimotor learning. *Proc Natl Acad Sci U S A*, 101(26):9839–9842.

[Landy and Kojima, 2001] Landy, M. S. and Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *J Opt Soc Am A Opt Image Sci Vis*, 18(9):2307–2320.

[Ma et al., 2006] Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci*, 9(11):1432–1438.

[Mackay, 1991] Mackay, D. (1991). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

[MacKay, 2003] MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

[Mansinghka et al., 2006] Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., and Griffiths, T. (2006). Structured priors for structure learning. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*.

[Nefian et al., 2002] Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphy, K. (2002). Dynamic bayesian networks for audio-visual speech recognition. *EURAISP Journal on Applied Signal Processing*, 11:1–15.

[Pasula et al., 2003] Pasula, H., Marthi, B., Milch, B., Russell, S., and Shpitser, I. (2003). Identity uncertainty and citation matching. In *Advances in Neural Information Processing 15 (NIPS 2002)*.

[Perez et al., 2004] Perez, P., Vermaak, J., and Blake, A. (2004). Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92(3):495–513.

[Petkos and Vijayakumar, 2007] Petkos, G. and Vijayakumar, S. (2007). Context estimation and learning control through latent variable extraction: From discrete to continuous contexts. In *Proc. IEEE International Conference on Robotics and Automation (ICRA '07),*.

[Pouget et al., 2003] Pouget, A., Dayan, P., and Zemel, R. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, 26:381–410.

[Rasmussen and Hager, 2001] Rasmussen, C. and Hager, G. D. (2001). Probabilistic data association methods for tracking complex visual objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:560–576.

[Recanzone, 2003] Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology*, 89:1078–1093.

[Roach et al., 2006] Roach, N. W., Heron, J., and McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proc Biol Sci*, 273(1598):2159–2168.

[Sato et al., 2007] Sato, Y., Toyoizumi, T., and Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Comput*, 19(12):3335–3355.

[Serby et al., 2004] Serby, D., Esther-Koller-Meier, and Gool, L. V. (2004). Probabilistic object tracking using multiple features. In *Proceedings of the 17th International Conference on Pattern Recognition*.

[Shams et al., 2000] Shams, L., Kamitani, Y., and Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, 408:788.

[Shams et al., 2001] Shams, L., Kamitani, Y., Thompson, S., and Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport*, 12(17):3849–3852.

[Shams et al., 2005] Shams, L., Ma, W. J., and Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport*, 16(17):1923–1927.

[Silva and Scheines, 2006] Silva, R. and Scheines, R. (2006). Bayesian learning of measurement and structural models. In *Proceedings of the 23rd International Conference on Machine Learning*.

[Siracusa and Fisher, 2007] Siracusa, M. R. and Fisher, J. W. (2007). Dynamic dependency tests: Analysis and applications to multi-modal data association. In *AIStats*.

[Slaney and Covell, 2000] Slaney, M. and Covell, M. (2000). Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Advances in Neural Information Processing Systems*.

[Stiefelhagen and Garofolo, 2007] Stiefelhagen, R. and Garofolo, J., editors (2007). *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*. Springer.

[Stone et al., 1999] Stone, L. D., Barlow, C. A., and Corwin, T. L. (1999). *Bayesian Multiple Target Tracking*. Artech House.

[Tenenbaum et al., 2006] Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends Cogn Sci*, 10(7):309–318.

[Treisman, 1996] Treisman, A. (1996). The binding problem. *Curr Opin Neurobiol*, 6(2):171–178.

[van Beers et al., 2002] van Beers, R. J., Wolpert, D. M., and Haggard, P. (2002). When feeling is more important than seeing in sensorimotor adaptation. *Current Biology*, 12:834–837.

[Vermaak et al., 2005] Vermaak, J., Godsill, S., and Perez, P. (2005). Monte carlo filtering for multi target tracking and data association. *Aerospace and Electronic Systems, IEEE Transactions on*, 41(1):309–332.

[Vijayakumar et al., 2001] Vijayakumar, S., Conradt, J., Shibata, T., and Schaal, S. (2001). Overt visual attention for a humanoid robot. In *Proc. International Conference on Intelligence in Robotics and Autonomous Systems*.

[Wallace et al., 2004] Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., and Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Exp Brain Res*, 158(2):252–258.

[Watkins et al., 2006] Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., and Rees, G. (2006). Sound alters activity in human v1 in association with illusory visual perception. *NeuroImage*, 31:1247–56.

[Williams et al., 2006] Williams, C. K. I., Quinn, J., and McIntosh, N. (2006). Factorial switching kalman filters for condition monitoring in neonatal intensive care. In Weiss, Y., Schoelkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*. MIT Press.

[Williams and Titsias, 2004] Williams, C. K. I. and Titsias, M. K. (2004). Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Comput*, 16(5):1039–1062.

[Witten and Knudsen, 2005] Witten, I. B. and Knudsen, E. I. (2005). Why seeing is believing: merging auditory and visual worlds. *Neuron*, 48(3):489–496.

[Yuille and Kersten, 2006] Yuille, A. and Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends Cogn Sci*, 10(7):301–308.

[Zhang et al., 2008] Zhang, C., Rui, Y., Crawford, J., and wei He, L. (2008). An automated end-to-end lecture capture and broadcast system. *ACM Transacations on Multimedia Computing, Communications and Applications (in press)*.